TÜBİTAK

2004-609

TÜRKİYE BİLİMSEL VE TEKNOLOJİK ARAŞTIRMA KURUMU

THE SCIENTIFIC AND TECHNOLOGICAL RESEARCH COUNCIL OF TURKEY

Elektrik, Elektronik ve Enformatik Araştırma Grubu

Electrical, Electronical and Informatics Research Group

COST 276 "Tümleşik çoklu ortam iletişimi için bilgi idaresi" (Information and Knowledge Management for Integrated Media Communication Systems)

PROJE NO: 101E036

DOÇ. DR. GÖZDE BOZDAĞI AKAR

Mart 2006

ANKARA

ÖNSÖZ

Açılış toplantısını Temmuz 2001 tarihinde yapmış COST 276 aksiyonu kendi içinde yapılanma ve belli alt konulara odaklanma süreci geçirmiş ve COST 276 aksiyonunun içerdiği geniş başlık altında 4 adet çalışma grubu (WG) oluşturulmuştur. O.D.T.Ü. Elektrik-Elektronik Mühendisliği Bölümü projeye dahil olduğu 1. 8. 2002 tarihinden itibaren aşağıdaki konulara yoğunlaşmış ve bu konularda araştırmalarını sürdürmüştür.

(a) çoğulortam bilgilerinin yönetimi için otomatik endeksleme metodları geliştirmek (WG1),

(b) dağınık bir çoğulortam bilgi yönetim/idare sistemi mimarisi gerçekleştirmek (WG2),

(c) gezgin ve ağ tabanlı sistemler için çoğulortam bilgi alışverişini sağlayacak algoritmalar oluşturmak (WG4).

2003 yılında COST 276 yönetim kurulu kararıyla yapılan değişiklik sonucu, aksiyon kapsamına görüntü ve video içine bilgi saklama ve damgalama konusu eklendikden sonra bu konuda da katkıda bulunulmuştur.

Bu proje ODTÜ Elektrik Elektronik Mühendisliği Bölümünde gerçekleştirilmiş olup TÜBİTAK EEEAG tarafından desteklenmiştir.

İÇİNDEKİLER

ŞEKİL LİSTESİ

ÖZ

Günümüzde çoğulortamlı teknolojilerin, iletişim ve yayıncılık sektörünün hızlı bir şekilde yakınsamasına şahit olmaktayız. Etkileşimli televizyon, ısmarlama video (video-on-demand), sayısal kayıt cihazları bu yakınsamanın ortaya çıkardığı önemli ürünler olarak bilinmektedir. Bu ürünlerin geliştirilmesi ve amaçlarına uygun hizmet edebilmesi için önemli olan unsurlar kullanıcı servislerinin verimli teslimi, otomatik içerik işleme, kişiselleştirme olarak sayılabilir. Bunların gerçekleştirilebilmesi için bu proje kapsamında 4 ana madde üzerinde yoğunlaşılmıştır :

- İçerik ve bilgi içeren öteveri (metadata) tanımları: Kullanıcılar için kişiselleştirilmiş içerik seçimi, bu içeriğin uygun bir biçimde tanımlanmasını (ör: öteveri kullanarak) gerektirmektedir. Günümüzde MPEG-7 gibi içerik tanımlamaya yönelik çeşitli standartlar bulunmaktadır ve bu madde gerçekleştirilirken bu verilerden yararlanılmıştır. Aynı zamanda bu madde kapsamında içerik ve dizin teknikleri üzerinde de durulmuştur.
- Çoğulortamlı sistem teknolojileri: Çoğulortamlı verinin etkili bir biçimde iletimini içermektedir. Burada amaç varolan telli ve gezgin ağ yapısı kullanılarak verinin en az kayıpla ve en uygun fiyatla kullanıcı tarafına ulaştırılmasıdır.
- Damgalama:: Bu birim çoğulortam verilerinin güvenirliğinin sağlanması için tasarlanmaktadır.
- Çoğulortamlı veri yönetimi: Bu maddede çoğulortamlı sistemlerde bulunan veri çokluğu göz önünde bulundurularak değişik veri tipleri için uygun tanımlamaların geliştirilmesi hedeflenmiştir.

Anahtar Sözcükler:
Çoğulortam, veritabanı, gezgin iletişim, damgalama, veri yönetimi

ABSTRACT

Recently, we have been facing dramatic convergence of multimedia, hypermedia technologies, i.e. personal mobile terminals, television, video and computer technologies. It became apparent that interactive television, video-on demand, computer technology and massive storage technologies, coupled with the information and knowledge management, would drive the integration of hypermedia technologies even closer. In these converging worlds of telecommunications, broadcasting and the Internet, multimedia content management is therefore a key factor in promoting efficient delivery of end user services. To address these technologies, in this project we concentared on 4 topics:

- Content and knowledge metadata descriptors
- Multimedia systems and terminals
- Watermarking
- Multimedia content management

Keywords:
Multimedia, database, wireless communication, watermarking, data management

LİTERATÜR ÖZETİ

Bu projenin temelini çoklu ortam verilerinin "idare edilmesi" oluşturmaktadır [1]. Bu tip bilgi yoğunluğu içeren verileri idare etmenin en temel yolu endekslemeden geçmektedir [5,6]. Çokluortam bilgilerinin dünya üzerinde her geçen gün çoğalması bilgiyi verimli bir şekilde kullanmak için çok hızlı ve insan kullanımı gerektirmeyen endeksleme algoritmalarına duyulan ihtiyacı beraberinde getirmektedir. Saklanan çokluortam bilgilerinin erişiminin kolaylaştırılması ve bir eşgüdüm sağlanması amacıyla son yıllarda yapılan araştırmalar hızla çoğalmış ve ISO/IEC JTC1/SC29/WG11 MPEG tarafından bir standardizasyona gidilmesi öngörülmüştür [2,3,4]. Bu çalışmalar yıllardır süregelen görüntü analizi araştırmalarının insanlık için bir ürüne döndüğü bir alan olmaktadır [4]. Temel tanımlayıcı bilgilerin (şekil, renk, hareket, yazı, yüz [7], v.b.) [2] ve bu temel bilgilerin bir araya getirilmesiyle elde edilen anlamsal üst seviye bilgilerin [3,8,9] bir standard çevresinde toparlanmasıyla görsel bilginin paylaşımı ve taraması kolaylaşacaktır. MPEG-7 standardının önümüzdeki günlerde görsel bilginin kullanımı konusunda çok belirleyici bir alması beklenmektedir. Bu kapsamda grubumuz geçmişte ilgili standarda çeşitli katkıları olmuştur [8-20].

Aynı zamanda İnternet'in başarısı ile birlikte, kablosuz iletişim ve erişimin hızlı gelişimi, mobil/kablosuz çoklu ortam uygulamaları ve servislerinde yeni bir çağ başlatmıştır. İnternet, kablosuz ve çoklu ortamın birleşmesi, araştırmada ve geliştirmede, çoklu ortam içeriğinin İnternet ve mobil kablosuz dünyalar arasında rahatça taşınabilmesini sağlayan yeni bir paradigma yarattı. Kablosuz ağlardaki bant genişliğinin artması ile birlikte, mobil görsel telefonlar, kişisel sayısal asistanlar ve video akış gibi yeni erişim yetenekleri insanların günlük yaşamlarını etkilemektedir. Bu tür mobil ağ uygulamaları, İnternet içeriğine her an, her yerden, her türlü cihaz ile gelişmiş erişim yetenekleri sağlamaktadır. Ancak, bütünleşen teknolojiler tek başına ve koordineli olmayan çabaların sonuçlarıdır. Tasarım gerekleri ve kriterleri, diğerlerininkilere uymamaktadır. Böylece, farklı tip ağları bağlarken, çoğu zaman oldukça fazla miktarda iş yapılması gerekmektedir. Örneğin, İnternet güvenilir bağlantı merkezli aktarma mekanizmaları sağlayan TCP/IP protokolü

tabanlıdır. TCP kullanan uygulamalar, dünya genelindeki İnternet ağının değişken bant genişlik kapasitesinden etkilenmemektedir.

Öte yandan, çoklu ortam içeriğinin iletilmesi iki nokta arasındaki bağlantının kapasitesine ve kalitesine oldukça bağımlıdır. Bu yüzden, çoklu ortam verisini TCP kullanarak İnternet üzerinden taşımak zordur [21]. Bu problemi adresleyen bir takım çözümler vardır [22]. Kablosuz/mobil ağlar sahneye çıktığında, gürültülü kanalların doğası gereği başka problemler ortaya çıkmaktadır [24]. Problemler temel olarak iki kategoriye ayrılır: kodlama teknikleri ve aktarma mekanizmaları. Bazıları, çok yollu sönme, gölgeleme, semboller arası girişim, ve gürültüdür. Kablosuz/mobil ağların bahsedilen problemlerine çok sayıda çözüm önerilmektedir [25][26][27].

Aynı zamanda, bilgisayar mimarisindeki gelişmeler, görüntü cihazları, güç kaynakları ve silikon teknolojisi yeni bir boyut açmıştır [28]. Hesaplama cihazları artık cep büyüklüğündedir [29][30]. Bu cihazlar İnternet bağlantısı olmadan düşünülemezler. Şu günlerde, GSM ve wi-fi veya GSM, kızılötesi ve Bluetooth gibi birden çok ağ erişimine sahip olan cihazlar bile mevcuttur. Bu cihazların çoklu ortam yetenekleri masaüstü bilgisayarlarınkilerle kıyaslanabilecek güce erişmiştir. Çoklu ortam ağ servisleri artık küçük cihazların ulaşabileceği mesafededir.Mobile cihaz pazarı ve kablosuz/mobil ağların hızlı gelişimi, bizleri çoklu ortam ağ servislerini mobil dünyaya taşımayı düşünmeye zorlamaktadır. Çoğulortam verisinin saklanması ve iletimindeki gelişmeler beraberinde güvenlik sorununu getirmektedir. Bu amaçla son yıllarda sayısal damgalama üzerine yoğun çalışmalar yapılmaktadır. Bu pekçok sayıdaki algoritma arasında bazıları dayanıklılık, görünmezlik, işlem maliyeti gibi temel işaretleme gereklilikleri açısından daha üstün performans sergilemektedirler [31-36].
Bu proje kapsamında yukarıda belirtilen çoğulortam verisinin işlenmesi, endekslenmesi, iletimi ve de damgalanması konusunda araştırmalar yapılmıştır. Yapılan çalışmalara geçmeden önce bu konularla ilgili genel bir özet verilecektir.

Parametrik Hareket Tanımlayıcısı: Herhangi bir parmetrik hareket modeline göretanımlama yapar. Bu modeller yer değiştirme, döngüsel, affine, perspektif ve karesel modellerdir.

Çoğulortam verisinin iletimi

Sayısal video iletimi, video konferans, uzaktan eğitim ve ısmarlama video gibi var olan ve gelişmekte olan pek çok Internet uygulamasının önemli bir bileşenidir. Video verilerinin Internet üzerinden iletmek için yükleme ya da akislandirma yöntemlerinden biri kullanilabilir. Yükleme yöntemiyle, kullanici tüm videoyu kendi diskine yükledikten sonra oynatıma geçer. Akışlandırma yöntemiyle video gönderimi ve oynatımı paralel olarak gerçekleştirilir. Akışlandırma yöntemi, yükleme yönteminde yaşanan uzun süreli bekleme ve yüksek kapasiteli disk kullanımı gereksinimlerini ortadan kaldırdığı için daha uygun bir yöntemdir [46]. Akışlandırma yöntemi, doğasi gereği gerçek zamanlı iletişim gerektirdiği için internet üzerinden iletimde bazı problemlerle karşılaşılır. Bu problemlerin kaynakları aşağıdaki gibi sıralanabiliir.

Bant genişliği: Video verileri, diger veri türlerine oranla çok daha yüksek bant genişliğine ihtiyaç duyarlar.

Gecikme: Sürekli ve senkron video akışı için, uçtan uca gecikme değerlerinin belli bir üst sınırı aşmamasi gerekmektedir.

Kayıplar: Tekrar gönderme mümkün olmadığı durumlarda kayıp paketler, video kalitesinin bozulmasına yol açar.

Sonuç olarak, IP yönlendiricilerinin yetersiz bağlantı bant genişlikleri ve paket işleme hızlarının bir sonucu olarak oluşan ağ sıkışıklığı bu tür uygulamalar için ciddi bir sorun teşkil etmektedir. Ağ sıkışıklığına bağlı olarak gecikmeye uğrayan ve kaybolan paketler alıcıda çözülen video'nun kalitesini belirgin şekilde azaltmaktadır. Bununla birlikte, çokluortamlı uygulamalarda tercih edilen taşıma katmanı protokolü olan UDP'de (User Datagram Protocol/Kullanıcı Datagram Protokolü) herhangi bir sıkışıklık denetim mekanizması bulunmamaktadır. Bu soruna bir çözüm olarak, çözülen video kalitesini

Çoğulortam verisinin işlenmesi ve endekslenmesi

Bilgisayar yazılım ve donanımındaki son gelişmeler elektronik bilginin daha kolay bir şekilde üretilmesi, işlenmesi ve saklanmasını sağlamıştır. Elektronik bilgi başta sadece yazılı metinden ibaretken, giderek artan bir oranda grafik, imge, animasyon, video, ses ve diğer çoğul ortam verileri de bu kapsama dahil olmaktadır. Bu verilerin çoğuna , hızlı ağlar, arama makineleri ve gözatma araçları vasıtasıyla WEB üzerinden erişilebilmektedir. Fakat verilere ulaşma amaçlı yapılan sorguların çoğu istenilen verilerden daha çok ilgisiz verileri listelemektedir. Çoğul ortam verilerinde durum daha kötüdür. Geleneksel çoğul ortam erişim yöntemleri arama yapan kişinin verdiği anahtar sözcüklere dayanması nedeniyle, verimlilikten uzaktır. Bu sebebten dolayı sayısal görüntülerin içerik tabanlı erişimi veritabanı yönetiminde aktif bir araştırma konusudur [37-43].

Çoğulortam veritabanlarına veri kaydetme ve geri elde etme için izlenecek belirli basamaklar vardır. İlk aşamada üretilen veri analiz edilerek gerekli öznitelik bilgileri çıkarılır. Bu bilgiler dizinlenerek veri tabanına kaydedilir. Her hangi bir kullanıcı da bu sistemde bulmak istediği şeyi tanımlayan veya benzeyen örnek bir veri ile sistemden sorgular. Bu sistemdeki kilit noktalardan birisi bu verinin analizi sonucu ortaya çıkacak öznitelik bilgilerinin ne olacağıdır. Bu sorunun çözülmesi ve yaygın bir şekilde kullanılması için bir standart gereklidir. Dizinlenecek verinin sadece metinsel tabanlı olduğu kabul edilirse, o zaman metinsel verideki kelimeler o dile ait olan sözlüklerde bulunabilir. Aynı şekilde; metinsel verideki cümleler de o dilin gramer kuralları içersinde tanımlı olduğuna göre metin tabanlı arama kuralları konabilir. Eğer veri bir resim veya bir film ise nasıl bir gramer kuralına veya nasıl bir sözlüğe sahip olunacağı sorusu ortaya çıkmaktadır. Çoğul ortam veri içeriğinin otomatik bir yol bulunarak dizinlenmesi amacı ile MPEG-7 standartı oluşturulmuştur ve hala da geliştirilmeye devam edilmektedir [44,45] MPEG-7'de çeşitli araç tipleri bulunmaktadır. Bunlar:

Tanımlayıcılar (D): Tanımlayıcı bir veriye ait olan herhangi bir öznitelik bilgisini kendi üzerinde saklayan ve gösteren bir elemandır. Örnek olarak bir videoda yer alan çerçeveler arasındaki renk değişim dağılımı gösterilebilir.

Tanımlama Şemaları (DS) : Hem tanımlayıcılar ve hem de tanımlama şemaları arasında yer alan ilişkileri betimler.

*Tanımlama Belirleyici Dil (DDL):*Tanımlayıcıların ve tanımlama şemalarının yaratılmasına ve aralarındaki ilişkilerin kurulmasına olanak sağlayan dildir. Bu XML tabanlı geliştirilmiş bir dildir.

Sistem Araçları: Tanımlayıcıların depolanması, iletimi ve yönetimi ile ilgili geliştirilmiş araçlardır.

Aşağıda görsel bilginin yardımı ile elde edilen tanımlayıcı ve tanımlama şemaları üzerinde daha detaylı durulacaktır.

Görsel tanımlayıcılar görsel veri, grafikler, resimler ve videolardan oluşur. Görsel tanımlayıcılar dört anabaşlık altında sınıflandırılmıştır [44]: Renk tanımlayıcıları, doku tanımlayıcıları, şekil tanımlayıcıları ve hareket tanımlayıcıları. Ayrıca bunların dışında insan yüzü için geliştirilmiş bir yüz tanımlayıcısı da yer almaktadır [44].

Renk Tanımlayıcıları: Bu bölümde altı adet renk tanımlayıcısı kısaca açıklanacaktır.

Renk Uzayı Tanımlayıcısı: Renk uzayı seçimine olanak tanır. MPEG-7 içerisinde RGB, YCbCr, HSV, HMMD ve tek renkli renk uzayları kullanılır.

Baskın Renk Tanımlayıcısı: Olasılıksal verilerden faydalanarak baskın olan rengin belirlenmesine izin verir.

Ölçeklenebilir Renk Tanımlayıcısı: HSV renk uzayının histogramından Haar Dönüşümü kullanılarak tanımlanır.

Resimlerin Grupları veya Çerçevelerin Grupları Tanımlayıcısı: Ölçeklenebilir renk tanımlayıcısının grup resimlere ve video çerçeve gruplarına uygulanmış halidir.

Renk Yapısı Tanımlayıcısı: Renk histogramına dayanılarak çıkarılır. Amaç çerçeve bölgelerdeki renk dağılımlarını tanımlamaktır.

Renk Yerleşim Planı Tanımlayıcısı: Bir bölgede veya tüm resimdeki renklerin uzamsal(spatial) yapısı çıkarılmaya çalışılır. Temel olarak DCT katsayılarından elde edilir.

Doku Tanımlayıcıları: Bu bölümde üç adet doku tanımlayıcısı kısaca açıklanacaktır.

Türdeş Doku Tanımlayıcısı: Her bir bölümün ortalama enerjisini ve farklı frekanslardaki dağılımlarını kullanarak tanımlama yapar.

Doku Tarayan Tanımlayıcı: İnsan algılamasına benzer bir durumda çalışır. Dokunun düzgünlüğü ve yönelimi dikkate alınır.

Kenar Histogram Tanımlayıcısı: Her bir resimdeki her bir bölüm içinde yer alan kenarların histogramı tutularak doku bilgisi olarak saklanır. Bloklar için her bir resim 4x4 olarak bölütlenir.

Şekil Tanımlayıcıları: Bu bölümde üç adet şekil tanımlayıcısı kısaca açıkllanacaktır.

Alan Tabanlı Şekil Tanımlayıcısı: İki boyutlu bir nesnenin veya alanın piksel dağılımıtanımlanır. Şekil içersinde boşluklar olabilir.

Çevrit (contour-shape) Tabanlı Şekil Tanımlayıcısı: Bir nesnenin çevriti üzerinden tümşekil tanımlanmış olur. Bu nedenle şeklin içersinin tam olarak dolu olması gerekir.

Üç Boyut Şekil Tanımlayıcısı: Üç boyutlu nesnelerin çokgenler yardımı ile tanımlanması yapılır.

Hareket Tanımlayıcıları: Bu bölümde dört adet hareket tanımlayıcısı kısaca açıklanacaktır.

Hareket İşlekliği Tanımlayıcısı: Bir videoda yer alan hareketin hızı, yönü ve uzamsal şekliile bilgileri tanımlar.

Kamera Hareketi Tanımlayıcısı: Kameranın üç boyutlu uzaydaki hareketlerini tanımlar.

Hareket Gezingesi (trajectory) Tanımlayıcısı: Bir nesnenin belirli bir zaman içersindeuzaydaki yer değiştirme gezingesini tanımlar.

artırmak ve iyi bir ağ vatandaşı olmak için, uygulama seviyesinde sıkışıklık denetimi ve ağ uyarlanabilir kodlama algoritmaları uygulanabilir [47,48].

Sıkışıklık denetimi genel olarak pencere tabanlı ve hiz tabanlı algoritmalarca gerçekleştirilir. Hiz tabanlı algoritmalar akışlandırma hızını ağın sağladığı bant genişliği olanaklarına uyumlu olarak ayarladıkları için bu şekilde adlandırılırlar. Sıkışıklık denetimi algoıritmalarına örnek olarak Busse-Deffner-Schulzrinne [49], DAA [50], LDA+ [51] ve TLFC [52] verilebilir. Bu algoritmalardan iki tanesi (Busse-Deffner-Schulzrinne ve DAA) belli bir oranda paket kaybına izin verirken diğer ikisi (LDA+ ve TLFC) sıfırdan büyük paket kaybını ağ sıkışıklığı olarak kabul etmektedir. Bu algoritmaların hepsi gerçek zamanlı video iletimi için RTP [53] (Real-time Transport Protocol/Gerçek Zamanlı İletim Protokolü), gönderilen video paketlerinin durumuyla ilgili geri besleme almak için de RTCP [53] (Real-time Transport Control Protocol/Gerçek Zamanlı İletim Denetim Protokolü) kullanmaktadırlar. Video alıcısı gönderilen paketlerle ilgili bazı bilgileri içeren (paket kayıp oranı gibi) bir RTCP raporu hazırlayarak video göndericisine gönderir. Gönderici, bu rapordaki bilgilere dayanarak ağın sıkışıklık durumunu belirleyip gönderilen video'nun bit hızını buna uygun şekilde artırır veya azaltır. İki RTCP alıcı raporu arasındaki süre en az beş saniye olduğu için bu algoritmalar ağ şartlarına anlık olarak uyum sağlamayı hedeflememektedirler. Bunun yerine ağ şartlarına ortalama olarak bir uyum sağlamaya çalışmaktadırlar.

Video kalitesini arttırmak için daha önce bahsedildiği gibi diğer bir çözüm de ağ koşulundaki değişimlere uyum sağlayan ve en iyi uçtan uca iletim performansını sağlayan ağ uyarlanabilir video iletim yöntemleridir. Yüksek gecikmeye sahip kanallarda genelde *önceden hata düzeltimi* (FEC) ve *çoğul tanımlı* (MD) kodlama kullanılmaktadır. Fakat bahsedilen kodlamalardaki artıklık(redundancy) yüzünden hız-bozunum performansında kayıp oluşmaktadır, bu sebeple daha yüksek performans sağlanabilen akıllı otomatik tekrarlama isteği (ARQ) algoritmalarına ilgi artmaktadır. Zakhor [54]'de paketlerin son(deadline)'larına göre heuristic tabanlı bir önceliklendirme algoritması önermiştir. Buna karşın [55]'de Chou hız-bozunuma göre en iyi (RaDiO) gönderim yönteminin hesaplanmasını göstermiştir. Bu algoritma ile hız-bozunum, kanal istatistiği,

paket sonu ve iletim geçmişi bilgilerini kullanarak hesaplanan bir Lagrange maliyet (cost) fonksiyonununu enküçülten paket iletim politikaları bulunmuştur. Oldukça popüler olan RaDiO iletimi fikri [56]'da yol değişkenliği (path diversity) senaryosuna, [57]'de ise artan artıklık (IR) olarak da bilinen bir hibrid FEC/ARQ iletimi senaryosuna uyarlanmıştır. RaDiO iletim algoritmasının performansı sistemin gözlenebilirliğine bağlıdır ve bu bilgi alındı (ACK) mesajları ile sağlanmaktadır. Son olarak ise [58]'de gönderen tarafında, alıcının durumu bir olasılık dağılımına göre tahmin edilip, en iyi politikalar POMDP modeli kullanarak hesaplanmıştır.

Damgalama

Sayısal damgalama ses, imge, ve video gibi sayısal bilgilerin içerisine görsel/işitsel olarak algılanmayacak, silinmeyecek, değiştirilmeyecek şekilde sahiplik haklarını temsil eden sayısal bilginin yerleştirilmesidir. Sayısal damganın amaçları arasında ses, imge ve video gibi sayısal ürünlerin üreticilerini belirlemek, kopyalarını izlemek, çoğaltım yetkisini sınırlamak gibi işlemler sayılabilir. Bu zamana kadar sunulan farklı damgalama teknikleri geniş olarak zaman bölgesinde ve frekans bölgesinde sunulan damgalama teknikleri diye iki grupta toplanabilir [31-36]. Zaman bölgesinde gerçekleştirilen imge damgalama yöntemlerinde, damgalanacak imge ile damga bilgisinin piksel değerlerinin değiştirilmesi neticesinde damgalama işlemi gerçekleştirilir. Ancak zaman bölgesinde gerçekleştirilen damgalama yöntemleri kolay ve az işlem gerektirmesine rağmen, damga bilgisinin imgenin her tarafına yayılmadığından damganın yok edilmesi veya ortaya çıkarılması maksadıyla yapılan saldırılara bu tür yöntemler dayanıksızdır. Bunun yanında frekans bölgesinde gerçekleştirilen damgalama tekniklerinde damga genelde düşük ve orta frekans katsayılarına yerleştirilir ve değişimler imgeyi kapsar. Böylece damgayı çıkartmak veya yok etmek için yapılan saldırılarda frekans bölgesinde gerçekleştirilen damgalama teknikleri zaman bölgesinde gerçekleştirilen tekniklere göre daha fazla dayanıklıdır. Bu sınıflamanın haricinde sınıflandırma genel ve özel damgalama şeklinde de yapılmaktadır. Bilgi saklamanın ilk dönemlerinde, gözü kapalı algılama (genel

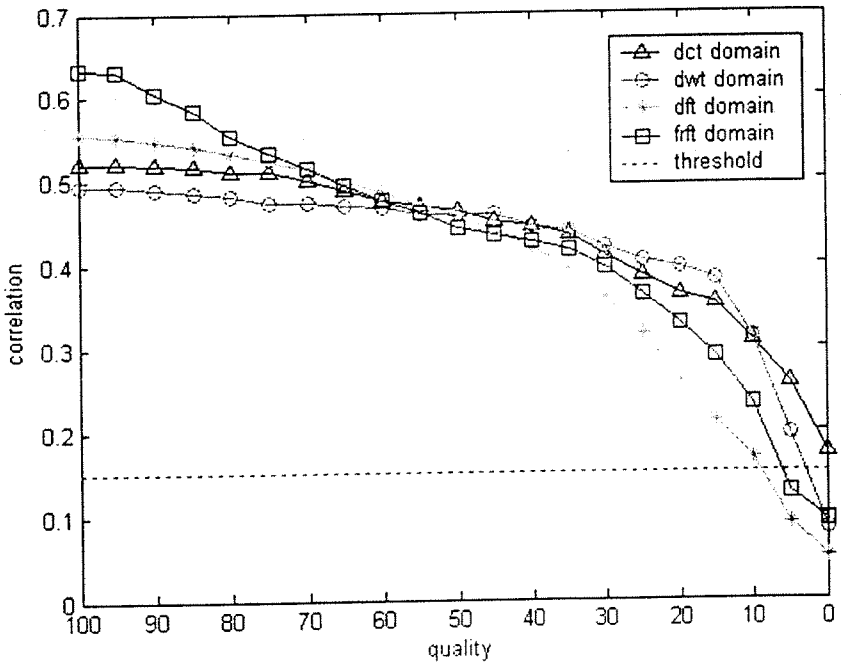
damgalama) üzerinde, bildirilmiş algılamaya (özel damgalama) kıyasla çok daha az durulmuştur. Bahsedilen ikinci yaklaşım, ilk bakışta daha kolay gibi gözükse de bilişim kuramını kullanan bazı son gelişmeler [59], gözü kapalı algılamaya olan ilgiyi artırmıştır. Costa'nın ünlü makalesi [60], bu konuda yeni yaklaşımlara yol açmıştır. Costa bu çalışmasında, ek bilginin sadece kodlayıcıda olduğu iletişim sistemi ile kodlayıcı ve kodçözücüde birlikte bulunduğu sistemin kapasitelerinin aynı olduğunu göstermiştir [60]. Bu sayede, aynı kapasiteye sahip özel ve genel damgalama yöntemlerinin tasarlanabileceği gösterilmiş olmuştur. Enformatik kuramı araçlarından ilham alan pek çok genel damgalama algoritması da geliştirilmiştir [61-64]. Chen [61,62], Nicemleme Dizin Modülasyonunu (QIM) yaklaşımını bilgi saklamanın genel bir sınıfı olarak sunmuştur. QIM bilgiyi örtü nesnesinin içine nicemleme kullanarak yerleştirir. Farklı bilgi için değişik nicemleyiciler kullanılır. Kodçözücüde, kodlayıcıdaki QIM yapısı kullanılarak saklanan bilgiler çıkarılır ve bu noktada özgün sinyale gerek yoktur. Bunun dışında nicemleme kullanan daha farklı yöntemler de önerilmiştir [63,64]. Biraz farklı bir yaklaşımla, kaynak kodlama ile kanal kodlama arasındaki benzerliği gözlemleyen Chou [65], nicemlemeyi kafes kodlamalı modülasyondan sonra kullanmıştır.

YAPILAN ÇALIŞMALAR

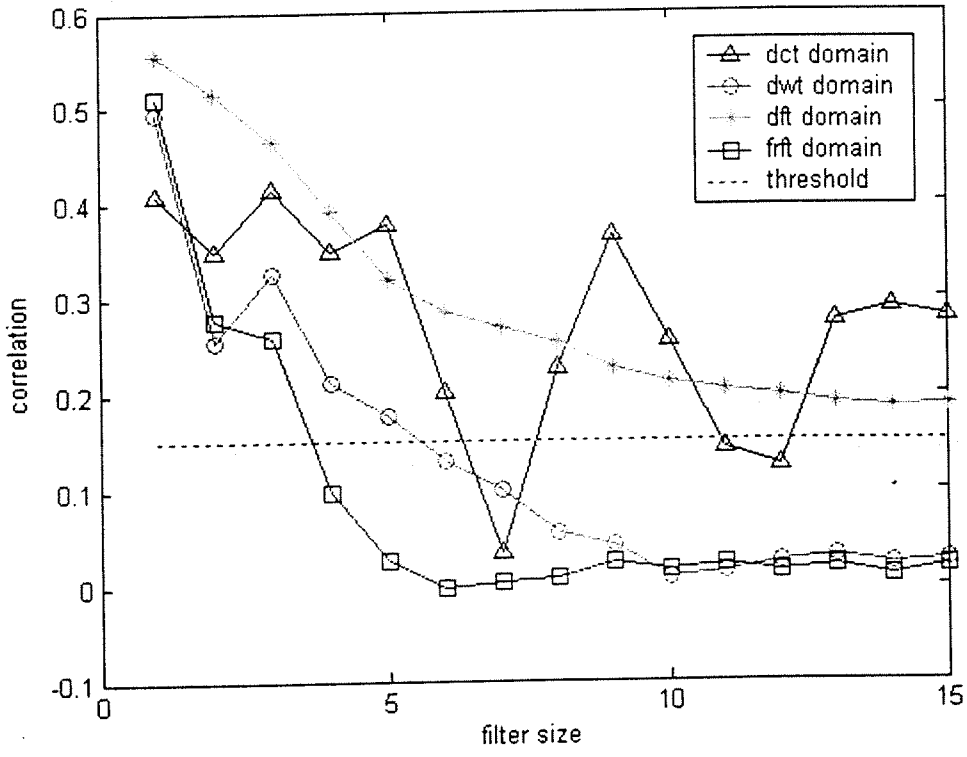
Bu proje kapsamında yapılan çalışmalar bu bölümde özet olarak verilecektir. Daha detaylı açıklamalar ise ekde verilen yayınlarda bulunabilir. Sadece tez olarak sonuçlanan çalışmalar için ise daha açıklayıcı bilgi ve sonuçlar verilmiştir.

Bilgi Saklama

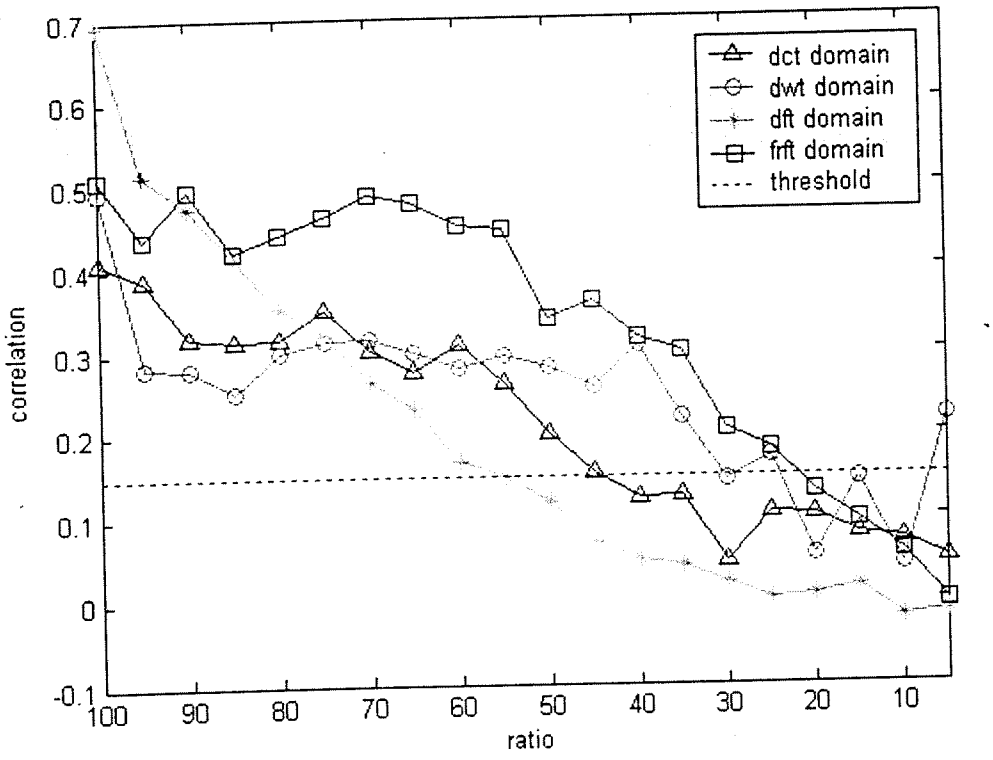
Bu konu kapsamında varolan tekniklerin kıyaslanması yapılmış ve de yeni algoritmalar önerilmiştir. Öncelikle hem uzamsal hem de frekans bölgelerinde yapılan çeşitli damgalama algoritmaların karşılaştırıldığı ve de stirmark'da bulunan tüm saldırı ve bozulmalara karşı testlerin gerçeklendiği bir çalışma yapılmıştır [66]. Daha sonra bu çalışma uzamsal ve frekans bölgelerini birleştiren kesirli Fourier dönüşümü kullanılarak damganın imgeye gömülmesi yönünde bir ilerletilmiştir. İmgenin geometrik saldırılara karşı dayanıklılık kazanması için damgalı imgeye Fourier dönüşüm uzayında, yine farkedilmeyen bir şablon eklenmiştir. Şablon özünde bilgi barındırmaz ama imge üzerine uygulanan geometrik saldırıların belirlenmesinde kullanılır. İmgeye gömülü şablonun bulunmasıyla uygulanan geometrik dönüşüm hesaplanabilir ve bu dönüşüm tersine alınarak damganın çözülebilmesi sağlanır. Ayrıca değişik dönüşüm uzaylarının kullanıldığı damgalama algoritmalarının performansı da incelenmiştir. Bu algoritmalar damga gömmek için ayrık kosinüs dönüşüm uzayı, ayrık Fourier dönüşüm uzayı ve ayrık dalgacık dönüşüm uzayı kullanan algoritmalardır. Kesirli Fourier dönüşümü damgalama algoritmasının ve bu algoritmaların performansları çeşitli saldırılara ve bozunumlara karşı denenmiş ve dayanıklılıkları karşılaştırılmıştır [67].



Şekil 1: Sıkıştırma sonucu damgadaki bozulmalar.



Şekil 2: Ortalama filtre sonucu damgadaki bozulmalar.



Şekil 3: Kesme sonucu damgadaki bozulmalar.

Bu çalışmalara ek olarak nicemleme tabanlı yeni bir bilgi saklama yöntemi önerilmiştir. Önerilen yöntem, bilinen nicemleme tabanlı temel yöntemlerle karşılaştırılmıştır. Birörnek ve Gauss kaynaklar kullanılan deneylerde, değişik kanal gürültüleri için yöntemin başarımı incelenmiştir. Deneyler sonucunda, önerilen yöntemin diğer yöntemlerden kanal gürültüsünün çok olduğu durumlarda daha iyi sonuçlar verdiği gözlemlenmiştir. Yöntem, rasgele kaynaklar dışında imgelere de uygulanmış ve ümit verici sonuçlar alınmıştır [68,69].

Bu proje kapsamında aynı zamanda görüntüye ek olarak videoda da damgalama konusunda çalışmalar gerçekleştirdik. Bu çalışmalar güvenlik açısından olduğu gibi kodlama performansını arttıracak bazı teknikleri de içermektedir. Veri hızını çok

arttırmadan DCT katsayılarında yapılan değişikliklerle ardışık çerçeveler hakkında bilgilerin birbirlerinin içerisinde saklanmasıyla hata düzeltmesi sağlanmıştır [70,71,72].

Çoğulortam bilgilerinin yönetimi için otomatik endeksleme metodları geliştirmek (WG1)

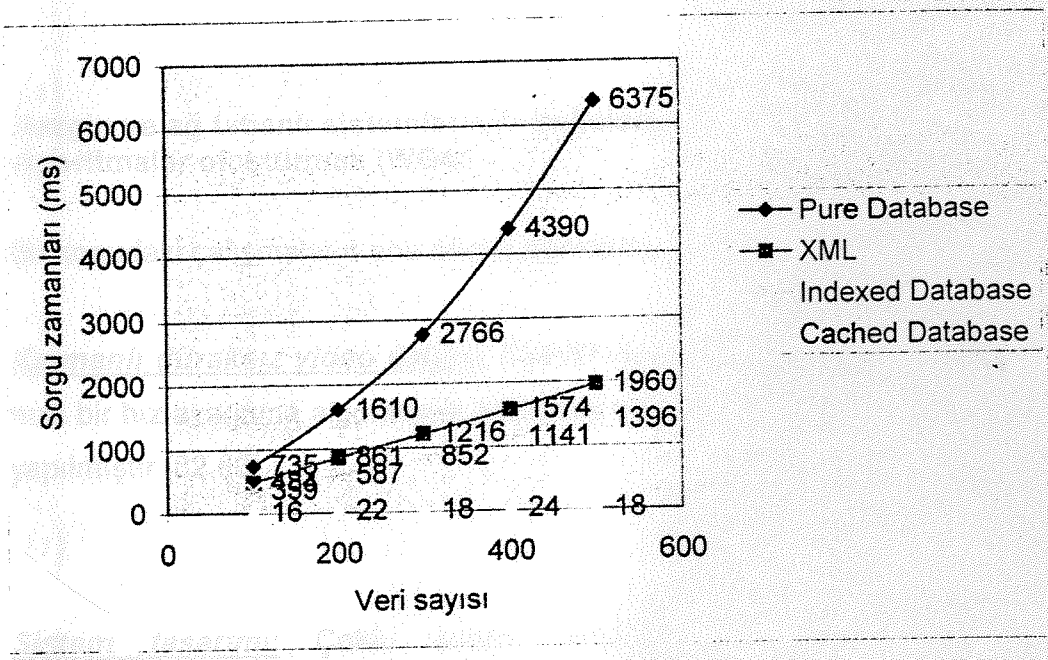
Bu konudaki çalışmalarımız çeşitli alt başlıklar halinde vurgulanacaktır.

İnsan yüzlerine dayalı endeksleme: İnsan yüzlerinden öznitelik vektörlerinin çıkarılması yönünde çalşmalarar kapsamaktadır. GABOR dalgacık dönüşümü kullanılarak elde edilen sonuçlar umut vericidir [73,74].

Anlamsal çıkarımlar: Video bilgisinden (TV, film) ses ve görüntü bilgileri kullanarak konuşmaların olduğu sahnelerin çıkarılması üst seviye bilgi özetlemesi için önemlidir. Bu konuda proje kapsamında videoların otomatik olarak endekslenmesiyle ilgili olarak Saklı Markov Model temelli bir yaklaşım üzerine çalışılmıştır [75,76]. .

Otomatik sınıflandırma: Sabit görüntüleri otomatik olarak görsel sınıflara bölecek sınıflandırıcı birleştiren özgün bir yöntem önerilmiştir [77,78].

Video için öte veri: MPEG-7 uyumlu olarak hazırlanmış XML tabanlı ve ilişkisel veritabanlarının performans analizlerinin yapıldığı bir tez gerçeklenmiştir [79]. Şekil 4'de çeşitli veri tabanları için sorgulama süreleri gösterilmektedir.



Şekil. 4: Sorgulama süreleri

Yazı bulma : Dalgacık dönüşümü tabanlı doku analizi, ayrılık tabanlı bölütleme ve renk uyumu göz önünde bulundurularak otomatik yazı bulma konusunda bir araştırma yapılmıştır [80].

Dağınık bir çoğulortam bilgi yönetim/idare sistemi mimarisi gerçekleştirmek (WG2)

Değişik eğitim uygulamarı için kullanılabilecek XML tabanlı verilerin depolanma metotları karşılaştırılmış ve yerli XML veritabanları için değişik sorgulama dilllerinin (Xpath ve XQuery) performansları arasında kıyaslama yapılmıştır. Kullanılanan depolama metotlarının ve sorgulama dillerinin geliştirlmesi hakkında çeşitli çalışmaların sürmesine karşın esas amaç, veritabanların performans ölçütlerinin anlaşılması ve yerli XML veritabanlarının veri depolama ve sorgulama açısından getirilerinin belirlenmesidir. Bu sebeplerden dolayı, ücretsiz bir yerli XML veritabanına belirli miltarda Öğrenme Nesnesi Yardımcı Verilieri (LOM) yüklenmiş ve de değişik sorgulama dilleri kullanılarak elde edilen sonuçlar değerlendirilmiştir [81].

Gezgin ve ağ tabanlı sistemler için çoğulortam bilgi alışverişini sağlayacak algoritmalar oluşturmak (WG4)

Bu konudaki çalışmalar 4 ana başlık altında toplanmıştır.

Katmanlı duraksız video iletimi: TCP/IP üzerinden katmanlı video iletimi için basit ve hızlı bir hız ayarlama algoritması önerilmiş, ns2 kullanılarak sistem performans deneyleri yapılmıştır [82,83].

Sistem tasarımı: Çoklu ortam verisine genel erişim sağlayan bir sistem gerçekleştirilmiştir. Çoklu ortam verisine erişmek, çoklu-ortam içeriğine her yerde olan bilgisayar ağları üzerinden farklı bilgisayar platformları kullanarak erişmek anlamına gelmektedir. Bahsedilen bilgisayar ağları hem kablolu hem kablosuz ağları; bahsedilen bilgisayar platformları kablolu kişisel bilgisayarları, taşınabilir bilgisayarları ve kişisel sayısal asistanları kapsamaktadır. Sistem istemci/sunucu mimarisi üzerine kurulmuştur. Görüntü verileri H.263 kodlanmış ve RTP üzerinden taşınmaktadır. Java Media Framework'ten yararlanılmış ve gereken durumlarda özel eklentiler ile yetenekleri genişletilmiştir. Gerçekleştirilen sistemde, H.263 kodlama için en uygun parametreler deneyler yardımı ile hesaplanmıştır [84].

Optimal video uyarlama: Çoğulortam verisinin sınırlı kaynaklara sahip mobil cihazlara gönderilirken optimal olarak adapte edilebilmesi irdelenmiştir. Son kullanıcının bir videoyu izlerken, videonun ona sunum şeklinden toplam tatmin olma miktarı fayda kuramı kullanılarak parametrik eğrilerle modellenmiştir. ITU-R BT.500-11 da belirtilen DSIS metodu kullanılarak elde edilen subjektif test değerleri, yukarıda bahsedilen modellerin test verilerine uymalarını sağlamak için kullanılmıştır. Kullanıcı terminalinin özellikleri (CPU gücü ve Ekran boyutları) sisteme parametre olarak girildikten sonra global optimizasyon algoritmaları(SA) kullanılarak kullanıcı tatmininin maksimum oldugu

video kodlama parametreleri (bit hızı, kare hızı ve piksel cinsinden çözünürlük) elde edilmektedir [85,86,87].

Aynı zamanda bu çalışma kapsamında Norveç Trondheim kentinde bulunan, NTNU Q2S merkeziyle de işbirliği içerisine girilmiştir. NTNU'da geliştirilen, veri işleme kapasitesi sınırlı mobil cihazlardan çoğul ortam verisine ulaşılabilmesini mümkün kılan teknoloji altyapısına ODTU EE de geliştirilen ve de çoğul ortam verisinin kullanıcı tercihleri ve cihaz yeteneklerine göre optimal olarak uyarlanmasını sağlayan algoritmaların entegre edilmesi yönünde çalışmalar yapılmıştır [88].

SONUÇLAR VE DEĞERLENDİRME

Bu proje kapsamında çoğulortam verilerinin idaresi, iletimi ve damgalanması yönünde çeşitli çalışmalar yapılmıştır. Bu çalışmaların sonucunda 7 Ms tezi, 14 konferans makalesi ve de 2 dergi makalaesi çıkarılmıştır. Proje kapsamında alınan kamera, PC ve telsiz cihazlar yapılan yayınlardaki deneyler için ve labaratuarımızda gerçekleştirdiğimiz diğer çalışamlar için verimli olarak kullanılmaktadır.

Aynı zamanda proje kapsamında doktora öğrencimiz Özgür Önür Norveç Trondheim kentinde bulunan, NTNU Q2S merkezinde geliştirilmiş olan UMATestBed ve MediaBase sistemlerinin incelenmesi ve de tarafımızdan geliştirilen algoritmaların sisteme entegre edilebilirliğinin incelenmesi amacıyla masrafları COST tarafından karşılanan 1 aylık bir STSM (short term scientific mission) gerçekleştirilmiştir.

Ayrıca 7. COST MC toplantısı ve çalıştayı 4-5 Kasım tarihlerinde Ankara'da düzenlenmiştir. Son derece başarılı geçen kongreyle ilgili tüm harcamalar COST tarafindan karşılanmıştır. Toplantıya devatli konuşmacı olarak Prof. Dr. Fernando Pereira katılarak MPEG-21 standardı hakkında bir konuşma yapmıştır. Bu sunuş aynı zamanda COST web sitesine de konmuştur.

COST 276 proje çalışanları olarak çıkarılması düşünülen kitapda da grubumuz elemanları aktif olarak yer alacaktır.

REFERANSLAR

1. Memorandum of Understanding for the implementation of a European Concerted Research Action designed as COST Action 276 "Information and Knowledge Management for Integrated Media Communication", Feb 2000
2. ISO/IEC JTC1/SC29/WG11/CD 15938-3 MPEG-7 Multimedia Content Description Interface – Part 3, Visual, January 2001.
3. ISO/IEC JTC1/SC29/WG11/CD 15938-5 MPEG-7 Multimedia Content Description Interface – Part 5, Multimedia Description Schemes, January 2001.
4. "Special Issue on MPEG-7 technology", Signal Processing, Image Communication, Elsevier, September 2000
5. R.M. Bolle, B. -L.Yeo and M.M.Yeung, ``Video Query : Research Directions," IBM Journal of Research and Development, vol. 42, pp.233--252, 1998.
6. Y. Rui, T.S. Huang and S.F. Chang, "Image Retrieval: Past, Present and Future", Journal of Visual Communications and Image Representations, vol. 10, pp. 1-23, 1999.
7. M.H.Yang, D.Kreigman and N.Ahuja, ``Detecting Faces in Images: A survey," to be published in IEEE Trans. on PAMI.
8. A.A.Alatan, A.N.Akansu and W.Wolf, ``Multi-modal Dialogue Scene Detection using Hidden Markov Models for Content-based Multimedia Indexing", to appear in Int. Journal on Multimedia Tools and Applications, Kluwer Ac., June 2001.
9. A.A.Alatan, A.N.Akansu and W.Wolf, ``Comparative Analysis of Hidden Markov Models for Multi-modal Dialogue Scene Indexing," Proceedings of ICASSP'2000, June 2000.
10. A.K. Peker, A. A. Alatan, A. N. Akansu, ``Low-level Motion Activity Features for Semantic Characterization of Video," Proceedings of IEEE ICME'2000 , NY., July 2000.
11. A.A.Alatan, ``Automatic Multimodal Dialogue Scene Indexing" submitted to ICIP'2001, October 2001

12. K. Peker, A.A.Alatan, A. N. Akansu, G. Bozdağı, "A Complementary Descriptor Proposal for Motion Activity Level", ISO/IEC JTC1/SC29/WG11 M4925, July 1999, Vancouver, CA
13. A.A.Alatan, A. Divakaran, Müfit Ferman, M. Tekalp "A proposal for Probability Model DS", ISO/IEC JTC1/SC29/WG11 M5244, October 1999, Melbourne, AU
14. K. Peker, A.A.Alatan and A. N. Akansu, "Comparison of Motion Activity Descriptors based on groundtruth", ISO/IEC JTC1/SC29/WG11 M5602, December 1999, Maui, US
15. A.A.Alatan and A. Divakaran, "A refinement to Probability Model DS", ISO/IEC JTC1/SC29/WG11 M5603, December 1999, Maui, US
16. J. R. Smith, Y-C. Chang, A.A.Alatan, A. Divakaran, T. Walker, "Report on Validation Experiments for MPEG-7 Probability Model DSs", ISO/IEC JTC1/SC29/WG11 M5735, March 2000, Noordwijkerhout, NL
17. T. S. Ingec, K. Ozbas, B. Tavli and G. Bozdagi, "MUSTER: Multi-platform SysTem for Efficient Retrieval from multimedia databases," ," CBMI'99, Oct. 1999, Toulousse.
18. R. de Queiroz, G. Bozdagi, Taha Sencar, ``A fast video segmentation technique for compressed video segmentation," SPIE'99, Jan. 1999, San Jose.
19. B. Kepenekci , F. B. Tek, G. Bozdagi, "Wavelet Based Face Recognition," Nesne Modelleme ve Oruntu Tanima Calistayi, May 2001, Istanbul.
20. G. Bozdağı, "Görüntü ve video veritabanlarına etkili erişim sistemi," EEEAG 100E106
21. Mahbub Hassan, Alfandika Nayandoro, Mohammed Atiquazzaman, "Internet Telephoy: Services, Technical Challenges, and Products", IEEE Communications Magazine, April 2000
22. DapengWu, Yiwei ThomsHou, Wenwu Zhu, "Streaming Video over the Internet: Approaches and Directions", pp. 282-300 IEEE Transactions On Circuits and Systems for Video Technology, Vol. 11, No. 3, Marc 2001.
23. H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, "RTP: A transport protocol for real-time applications" IETF Audio/Video Transport Working Group, January 1996, RFC 1889.

24. C. Wen Chen, R. I. Lagendijk, A. R. Reibman, W. Zu, "Inroduction to the Special Issue on Wireless Communication", IEEE Transactions On Circuits and Systems For Video Technology, Vol. 12, No. 6, June 2002
25. M. Van der Schaar, H. Radha, "Scalable Video Source Coding", IEEE Transactions On Circuits and Systems For Video Technology, Vol. 12, No. 6, June 2002
26. T.-C. Wang, H.-C. Fang, L.-G. Chen, "Low-Delay and Error-Robust Wireless Video Transmission for Video Communications", IEEE Tran-sactions On Circuits and Systems For Video Technology, Vol. 12, No. 12, December 2002
27. Q. Zhang, Z. Ji, W. Zhu, Y.-Q. Zhang, "Power-Minimized Bit Allocation for Video Communication Over Wireless Channels", IEEE Transactions On Circuits and Systems For Video Technology, Vol. 12, No. 6, June 2002
28. C. De Vleeschouwer, T. Nilsson, K. Denolf, and J. Bormans, "Algorithmic and Architectural Co-Design of a Motion Estimation Engine for Low-Power Video Devices", IEEE Transactions On Circuits and Systems For Video Technology, Vol. 12, No. 12, December 2002
29. PocketPC Magazine, http://www.pocketpcmag.com
30. Palm Products, Services, Company Information, http://www.palm.com
31. E. Koch and J. Zhao. "Toward robust and hidden image copyright labeling," in Proc. 1995 IEEE Workshop Nonlinear Signal and Image Processing, North Marmaras, Greece, June 20-22, 1995, pp. 452-455.
32. J. J. Quisquater, O. Bruyndonckx, and B. Macq, "Spatial method for copyright labeling of digital images," in Proc. 1995 IEEE Workshop Nonlinear Signal and Image Processing, North Marmaras, Greece, June 20-22, 1995, pp. 456-459.
33. Pitas and T. H. Kaskalis, "Applying signatures on digital images," in Proc. 1995 IEEE Workshop Nonlinear Signal and Image Processing, North Marmaras, Greece, June 20-22, 1995, pp. 460-463.
34. S. Craver, N. Memnon, B. L. Yeo, and M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks and implications," IEEE Trans. On Selected Areas of Communications, 16(4):573-586, 1998.

35. W. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding," Proc. SPIE, vol. 2420, pp. 40, 1995.
36. Ingemar J. Cox, Matt L. Miller, and Jeffrey A. Bloom, "Digital Watermarking," Morgan Kaufmann Publishers, 2002, pp. 21, Section 2.1.5.
37. Jorma Laaksonen, Markus Koskela, and Erkki Oja, "PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions," IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing, vol. 13, no. 4, pp. 841–853, July 2002.
38. Xiang Sean Zhou and Thomas S. Huang, "Relevance feedback for image retrieval: A comprehensive review," Multimedia Systems, vol. 8, no. 6, pp. 536–544, April 2003.
39. Markus Koskela and Jorma Laaksonen, "Using long-term learning to improve efficiency of content based image retrieval," in Proceedings of Third International Workshop on Pattern Recognition in Information Systems (PRIS 2003), Angers, France, April 2003, pp. 72–79.
40. Mats Sjoberg, Jorma Laaksonen, and Ville Viitaniemi, "Using image segments in PicSOM CBIR system," in Proceedings of 13th Scandinavian Conference on Image Analysis (SCIA 2003), Halmstad, Sweden, June/July 2003, pp. 1106–1113.
41. Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos, "Automatic image captioning," in Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, June 2004.
42. Jianping Fan, Yuli Gao, and Hangzai Luo, "Multilevel annotation of natural scenes using dominant image components and semantic concepts," in Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, Oct. 2004, pp. 540–547.
43. Mika Rummukainen, Jorma Laaksonen, and Markus Koskela, "An efficiency comparison of two contentbased image retrieval systems, GIFT and PicSOM," in Proceedings of International Conference on Image and Video Retrieval (CIVR 2003), Urbana, IL, USA, July 2003, pp. 500–509.

44. B. S. Manjunath, P. Salembier, T. Sikora, "Introduction to MPEG-7 Multimedia Content Description Interface", John Wiley & Sons Ltd., 2002.
45. ISO/IEC, "Information technology – Multimedia content description interface - Part 3: Visual," 15938-3:2002(E).
46. Kantarci A., Tunali T., "Design and Implementation of a Video on Demand System for the Internet", Packet Video 2000 Conference, 1-2 May 2000, Sardinia, Italy.
47. Wu D., Hou Y.T., Zhang Y.-Q.," Transporting Real-time Video over the Internet:Challenges and Approaches", Proceedings of the IEEE, VOL. 88, NO. 12, December , 2000.
48. Wu D., Hou Y. T., Zhu W., Zhang Y.-Q, Peha J. M., "Streaming Video over the Internet:Approaches and Directions", IEEE Transactions on Circuits and Systems for Video Technology, Vol 11, No.1, February, 2001.
49. Busse, I., Deffner, B., Schulzrinne, H., "Dynamic QoS Control of Multimedia Applications based on RTP," IEEE Computer Communications, January 1996.
50. Sisalem, D., "Fairness of Adaptive Multimedia Applications," IEEE International Conference on Communications, Volume 2, 1998.
51. Sisalem, D., Wolisz, A., "LDA+: A TCP-friendly Adaptation Scheme for Multimedia Communication," Proceedings of IEEE International Conference on Multimedia & EXPO 2000, Volume 3, July 2000.
52. Na, S., Ahn, J., "TCP-like Flow Control Algorithm for Real-Time Applications," Proceedings of IEEE International Conference on Networks, pages 99-104, 2000.
53. Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V., "RTP: A transport protocol for real-time applications," IETF Audio/Video Transport Working Group, January 1996, RFC 1889.
54. S. H. Kang and A. Zakhor, "Packet scheduling algorithm for wireless video streaming," Packet Video Workshop, Pittsburgh, April 2002.
55. P. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," Microsoft Research Tech. Rep., MSR-TR-2001-35, Feb. 2001.
56. J. Chakareski and B. Girod, "Rate-distortion optimized packet scheduling and routing for media streaming with path diversity," Proc. IEEE Data Compression Conference, Snowbird, UT, April 2003.

57. J. Chakareski and P. A. Chou, "Application layer error correction coding for rate-distortion optimized streaming to wireless clients, IEEE Transactions on Communications, vol.52, no.10, Oct. 2004.
58. D. Tian, X. Li, G. Al-Regib, Y. Altunbasak and J. R. Jackson, et.al., "Optimal packet scheduling for wireless video streaming with error-prone feedback," Proc. IEEE WCNC, Atlanta, Mar. 2004.
59. P. Moulin, "The role of information theory in watermarking and its application to image watermarking," invited paper, Signal Processing., vol. 81, no. 6, pp. 1121-1139, June 2001.
60. M. Costa, "Writing on dirty paper," IEEE Trans. Inform. Theory, vol.29, pp.439-441, May 1983.
61. B. Chen, Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems, Phd. Dissertation, MIT, 2000.
62. B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," IEEE Trans. Inform. Theory, vol. 47, no.4, May 2001.
63. J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between data hiding and distributed source coding," Invited Paper, Proc. 33rd Annual Asilomar conference on Signals, Systems, and Computers, Pacific Grove, CA, Nov. 1999.
64. J. Chou, S. S. Pradhan, L. El Ghaoui, and K. Ramchandran, "A robust optimization solution to the data hiding problem using distributed source coding principles," Proc. SPIE conference, San Jose, CA, Jan. 2000.
65. S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and Kanal coding and its extension to the side information case," IEEE Trans. Inform. Theory, vol. 49, no.5, May2003.

PROJE KAPSAMINDA YAPILAN YAYINLAR

66. Tolga Gökozan, Template Based Image Watermarking In The Fractional Fourier Domain, MSc Thesis, Ocak 2005, Danışman: G. B. Akar.

67. Salih Eren Balcı, Robust Watermarking of Images, MSc Thesis, Eylül 2003, Danışman: G. B. Akar.

68. E. Esen and A. A. Alatan, "Data Hiding Using Trellis Coded Quantization", IEEE ICIP 2004, 24-27 October 2004, Singapore.

69. E. Esen, A. Alatan, M. Aşkar, "Trellis Coded Quantization for Data Hiding", Eurocon 2003, Slovenya.

70. Ayhan Yılmaz, Robust video transmission using data hiding, Ayhan Yılmaz, MSc Thesis, 2003, Danışman: A. A. Alatan.

71. A. Yilmaz and A. A. Alatan, "Error Concealment of Video Sequences by Data Hiding", IEEE ICIP 2003, September, Barcelona, SPAIN.

72. A. Yilmaz, E. Esen and A. A. Alatan, "Combined Concealment, Synchronzation, and Error Detection Using Data Hiding" 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), April 2003, London, UK.

73. B. Kepenekci, F. Boray Tek, O. Çilingir, U. Sakarya, G. B. Akar, "GAYE: A Face Recognition System," SPIE/IST Electronic Imaging 2004, Jan. 2004, San Jose.

74. K. Messer , J. Kittler , M. Sadeghi , S. Marcel, C.Marcel , S. Bengio, F. Cardinaux, C. Sanderson , J. Czyz , L. Vandendorpe , S. Srisuk , M. Petrou , W. Kurutach , A. Kadyrov , R. Paredes , B. Kepenekci , F. B. Tek , G. B. Akar , F. Deravi , N. Mavity, " Face Verification Competition on the XM2VTS Database," Lecture Notes in Computer Science, vol. Volume 2688, pp. 964-974, 2003.

75. Yağız Yaşaroğlu, Multi-modal video summarization using HMMs for content-based multimedia,. MSc Thesis, 2003, Danışman: Aydın Alatan.

76. Y. Yasaroglu, A. Aydin Alatan, "Summarizing Video: Content, Features & Hmm Topologies", VLBV03.

77. M. Soysal, A. Aydin Alatan, "Combining MPEG-7 Based Visual Experts For Reaching Semantics",VLBV'03.

78. Medeni Soysal, Combining image features for semantic descriptions, MSc Thesis, 2003, Danışman: Aydın Alatan.

79. K. K. Güner, MPEG-7 Compliant ORDBMS Based Image Storage and Retrieval System, MS Thesis, 2004, Danışman: G. B. Akar.

80. S. Tekinalp, A. Alatan, "Utilization of Texture, Contrast and Color Homogeneity for Detecting and Recognizing Text from Video Frames", ICIP 2003.

81. E. O. Okman, G. B. Akar, "Evaluation of native XML databases for educational applications," COST 276 workshop, Oct. 2005.

82. E. Gurses, G. Bozdagi Akar, N. Akar, " Selective Frame Discarding for Video Streaming in TCP/IP Networks," Packet Video Workshop 2003, April 2003, Nantes, France.

83. E. Gurses, G. Bozdagi Akar, N. Akar, "Layered Video Streaming for Smooth Playout in TCP/IP Networks," Computer Networks, Volume 48, Issue 4 , 15 July 2005, Pages 489-501.

84. Ü. Ünal, A. Aksay, G. B. Akar, "An implementation of a wireless streaming system," COST 276 workshop, Oct. 2005.

85. O. D. Onur, Optimal Video Adaptation for Resource Constrained Mobile Devices based on Utility Theory, MSc Thesis, 2003, Danışman: A. A. Alata.n

86. O. D. Onur, A. A. Alatan, "Optimal Video Adaptation for Resource Constrained Mobile Devices based on Utility Theory", WIAMIS 2004.

87. O. D. Onur, A. A. Alatan, "Video Adaptation for Transmission Channels by Utility Modelling," ICME 2005.

88. O. D. Onur, P. Drege, A. Perkis, A. A. Alatan, R. Solberg, "Delivering Adaptive Content to Terminals Using Mpeg-21 Digital Items," COST 276 workshop, Oct. 2005.

Ekler

Proje kapsamında çıkan yayınlar.

DATA HIDING USING TRELLIS CODED QUANTIZATION[1]

Ersin Esen[2,3] and A. Aydın Alatan[2,3]
[2]Department of Electrical and Electronics Engineering, M.E.T.U.
[3]TÜBİTAK BİLTEN, Balgat 06531 Ankara TURKEY

ABSTRACT

Information theoretic tools lead to the design and analysis of new blind data hiding methods. A novel quantization-based blind method, which uses trellis coded quantization, is proposed in this manuscript. The redundancy in initial state selection during trellis coded quantization is exploited to hide information as the index of this initial state. This index is recovered at the receiver by Viterbi decoding after comparison with all initial states. The performance of the proposed method is compared against other well-known approaches via simulations and promising results are obtained. Based on these results, the proposed method can be preferred in certain applications with high distortion attacks.

1. INTRODUCTION

In the early stages of data hiding, mostly *informed detection (private watermarking)* has been examined, compared to *blind detection (public watermarking)*. Although, the former problem intuitively appears to be simpler than the latter, some recent developments based on information theoretic tools [1], caused blind detection problem to attract more attention. Costa's paper [2] is one of the most remarkable works, which results in the analysis and design of novel blind detection methods. Costa shows the equality between the capacity of a communication system with side information available only to encoder and that of the system with side information available to both encoder and decoder [2]. Hence, it is possible to devise public watermarking systems achieving the same capacity with private systems.

Inspired by the information theoretic tools, many public watermarking algorithms are developed and analyzed [3-6]. Chen et.al. [3,4] has introduced Quantization Index Modulation (QIM), as a generic class of data hiding methods. QIM embeds data into cover content using quantization. Different quantizers are used for different watermarks or data; i.e. quantizers are modulated according to the data. At the decoder, the hidden data is decoded using the knowledge of the QIM structure employed at the encoder; hence, the original content is not required. Other public data hiding method based on quantization are proposed in [5,6]. In a slightly different approach, Chou et.al. [7] use quantization after trellis coded modulation as a consequence of the observation of the duality between source coding and channel coding with side information.

In this paper, we propose yet another quantization-based public data hiding method. The method is mainly based on Trellis Coded Quantization (TCQ), which is described briefly in Section 2. The proposed method is presented in Section 3, as well as two other quantization based methods. An algorithm that utilizes the proposed method to embed hidden information into the images is also described in Section 4. Following the simulations in Section 5, the concluding remarks are given in Section 6.

2. TRELLIS CODED QUANTIZATION

TCQ [8] can be considered as a special case of trellis coding. The main ideas behind TCQ are due to trellis coded modulation (TCM) [9]. TCQ employs a set of trellises and set partitioning ideas of TCM in order to achieve better distortion performance with low complexity.

TCQ uses a trellis and an associated codebook. A sample trellis of 4 states is shown in Figure 1, in accordance with its finite state machine and the codebook. In this figure, u is the input bit of the finite state machine, s_1, s_0 are the states, and o_1, o_2 are the outputs of this machine. The smooth lines in the trellis correspond to $u=0$ and the dashed lines correspond to $u=1$. Each branch of the trellis is associated with a subset D_i of the codebook, whose index i is determined by the output of the finite state machine. In order to design the codebook of TCQ, a scalar codebook C of size $2^{R+\acute{R}}$ is taken. R is the encoding rate and $\acute{R}$ represents the number of bits that specify the particular codeword in the selected subset. For the sample system in Figure 1, R is 2 bits per sample (bps), whereas $\acute{R}$

[1] This work is partially supported by TÜBİTAK under project EEEAG 101E007

is equal to 1. For an encoding rate of 2 bps, C is twice larger than the corresponding scalar quantizer. After C is determined, it is partitioned into $2^{\acute{R}+1}$ subsets, each of which has $2^{R-\acute{R}}$ codewords.

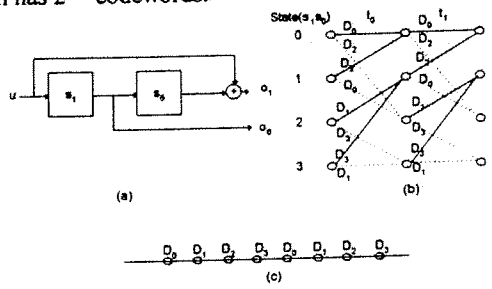


Fig. 1. A typical 4-state TCQ structure.

In order to quantize an input sequence s of length m, a trellis of m stages is used. Viterbi algorithm is used to find the closest path to s among all possible trellis paths. $m(R-\acute{R})$ bits specify the trellis path and $m\acute{R}$ bits specify the codeword in the selected subset. Although initial state selection is completely arbitrary, this state is generally selected as 0, since for long data sequences ($m>>log_2N$, where N is the number of states) the effect of the initial state on Mean Square Error (MSE) is negligible [8].

3. QUANTIZATION-BASED DATA HIDING

In addition to the proposed method, two other well-known methods, Dither Modulation (DM) and TCQ-Path Selection (TCQ-PS), are also briefly described in this section.

3.1.Proposed Data Hiding Method : TCQ-IS

The proposed method makes use of TCQ while considering the effects of the initial state (IS) on MSE. Since initial state selection is arbitrary in TCQ, one can embed information into the content by enforcing the selection of the initial state, accordingly.

After enforcing the selection of the initial state according to the data to be hidden, the closest trellis path is found using Viterbi algorithm, as in conventional TCQ.

Figure 2 shows a sample embedding process for TCQ-IS. In this illustrated example, $N=4$, $m=3$, $R=2$, $\acute{R}=1$. Since N is 4, one can hide 2 bits into the m input samples. Assuming, the data to hide is arbitrarily selected as $w=\{0,1\}$, state 1 is chosen as the initial state. The corresponding closest path found by the Viterbi algorithm is shown with thicker branches.

The total number of bits that can be hidden by TCQ-IS is log_2N, which depends only on the number of states in the trellis. Embedding rate of TCQ-IS is log_2N/m bps. For a given m, one should use trellises with more states in order to hide more data.

In the decoding stage, all N states are considered as the candidate initial states. For each candidate initial state Viterbi algorithm is executed for the received input vector. The computed MSE values are stored and the initial state with minimum MSE gives the desired embedded data, as the index of the corresponding state.

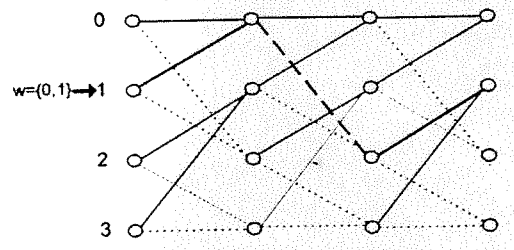


Fig. 2. A TCQ-IS sample.

It should be noted that the embedding distortion due to information hiding is mainly determined by the bin widths of the quantizer (Fig. 1), similar to other methods.

3.2. Dither Modulation (DM)

DM is a member of the QIM family. DM employs a quantizer set in which the quantization cells and the reconstruction points are the shifted versions of any other quantizer in the set [3]. The dither vector d is modulated with the watermark w. Then, using the corresponding dither vector for a given w, the embedding function is,

$$x(s,w) = q(x + d(w)) - d(w),$$

where q is the selected base quantizer. A uniform quantizer with step size Δ is used as the base quantizer. The dither vectors are $-\Delta/4$ and $+\Delta/4$ for $w=0$ and $w=1$, respectively. The hidden data is decoded by computing the distance between the reconstruction levels of each quantizer in the set. The index of the quantizer with the minimum distance gives the decoded data. According to the desired embedding rate, input data s can be used individually or sequentially in conjunction with majority voting in the decoding stage.

3.3.TCQ-PS

TCQ-PS is based on the TCM-TCQ scheme described in [5,6]. Using the same trellis structure in TCQ-IS, the coset selection in which the input signal s is to be quantized reduces to the trellis path determination according to the data w. Once the path is determined, s is quantized using the corresponding D_i. In that sense, TCQ-PS is also a member of QIM family. However, in this case, the quantizers are selected according to a trellis. Figure 3 shows a sample for TCQ-PS. The trellis path is determined with respect to the data $w=\{0,1,1\}$ starting always from the initial state 0. The data is quantized using the associated subsets of the branches in the order dictated by the trellis. In this scheme, embedding rate is 1 bps. At the decoding side, Viterbi algorithm is used to quantize the received

signal using the trellis structure. Once the best path is found, the data is decoded starting from the initial state 0.

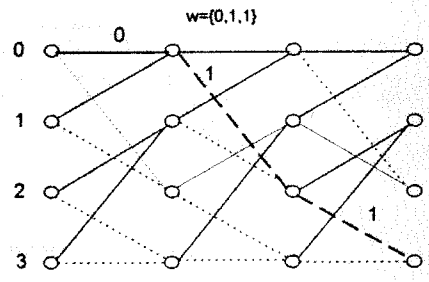


Fig. 3. A TCQ-PS sample.

4. IMAGE DATA HIDING BY TCQ-IS

TCQ-IS described in Section 3.1 is applied to images after transform domain conversion. For this purpose, the image is partitioned into blocks and for each block, Discrete Fourier Transform (DFT) is performed. The coefficients, for which TCQ-IS will be applied, are selected from the middle frequency band, as shown in Figure 4. The low frequency coefficients are not included, since modifications on these coefficients will be more visible. The high frequency band is also not considered, since the coefficients in this band are expected not to survive compression. The magnitudes of the selected mid-frequency DFT coefficients are fed into TCQ-IS. Resulting quantized values are replaced with the originals and finally, the cover image is obtained.

The decoder uses the same set of coefficients and extracts the hidden bits after the Viterbi decoding. Figure 4 shows a typical cover image (gray level *Lena* of size 512x512) with embedding distortion (PSNR between original and cover image) of 43dB.

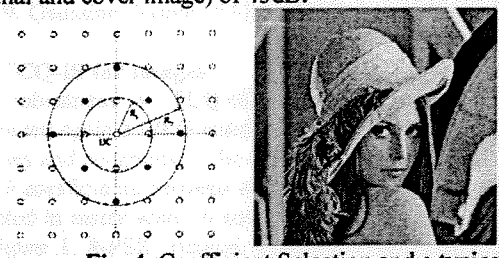


Fig. 4. Coefficient Selection and a typical example.

5. SIMULATIONS

5.1. Effect of the Sequence Length on TCQ-IS

The robustness performance of TCQ-IS against data sequence length is shown in Figure 5. The robustness is measured by the relation between the probabilities of erroneous bit decoding versus the watermark-to-noise ratio (WNR). Figure 5 displays the robustness of TCQ-IS with different lengths for Gaussian input source with zero mean, unit variance against a Gaussian channel noise of zero mean, 0.5 variance. As apparent from Figure 5, increasing data length improves the performance. However, beyond a limit, the improvement becomes indistinguishable. The reason for such a performance should be due to the fundamental problem (i.e. $m>>log_2N$) for detecting the initial state, as the length increases.

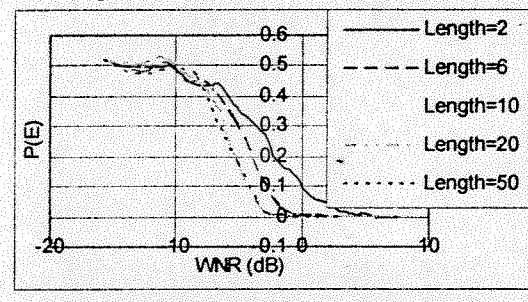


Fig. 5. TCQ-IS for different sequence lengths.

5.2. Comparison between TCQ-IS, DM, and TCQ-PS

The robustness performances of the 3 methods described in Section 3 are examined in terms of WNR and the probability of error, P(E). The channel noise power is kept constant, while the watermark power is varied. The probability of error is computed as the ratio of erroneously decoded data bits to the number of total data bits. The final values are computed by averaging the results for 500 random experiments. Two sources are used to embed data: a uniform source in the range (-1,1) and a Gaussian source with zero mean and unit variance. The TCQ structure in Figure 1 is utilized for both TCQ-IS and TCQ-PS during these simulations. The codebook is chosen so that it covers all input data range for the uniform source. For Gaussian source, the selected range contains most of the signal energy. Input data length is selected as 10. In order to equate the data embedding rate, DM uses 1 sample to hide 1 bit and TCQ-IS operates on input samples in lengths of 2. The robustness of the methods is plotted with respect to various Gaussian channel noises in Figures 6-9 .

The results indicate that, for all low-noise cases, TCQ-IS and DM performs similar, which is better than TCQ-PS. On the other hand, for high noise case, TCQ-IS has the best performance among the three methods.

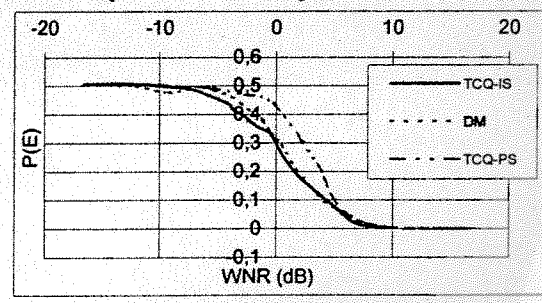


Fig.6. Uniform Source (-1,1) and Gaussian (0,0.5) Channel

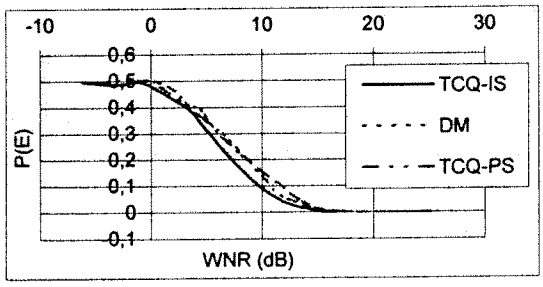


Fig.7. Uniform Source (-1,1) and Gaussian (0,1) Channel.

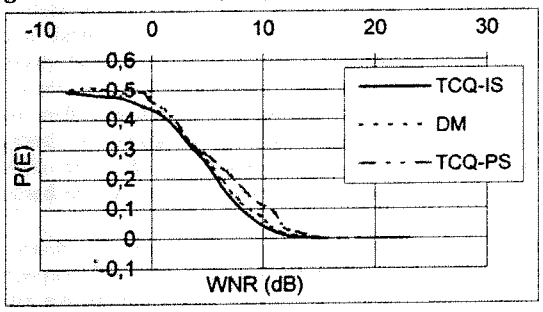


Fig 8. Gaussian Source (0,1) and Gaussian (0,1) Channel.

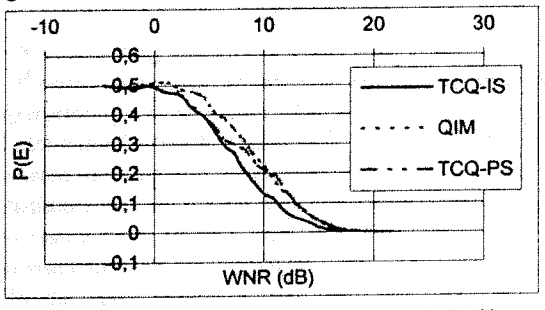


Fig.9. Gaussian Source (0,1) and Gaussian (0,2) Channel.

5.2. TCQ-IS for Images

The robustness of TCQ-IS for image data hiding is computed against JPEG compression attack using different trellises and codebooks. The block size is taken as 16. The first 6 coefficients between the circles of radii 3 and 2 are selected in raster scan. In addition to 4-state trellis shown in Figure 1, 8-PSK trellises with 8 and 16 states [9] are also employed. The results are shown in Figure 10, as the embedding distortion versus probability of error against JPEG-80 compression attack. It is apparent that as the trellis structure becomes dense, the robustness increases. It is also observed that even if the state numbers differ two PSK schemes behave similarly, since they share the same codebook.

6. CONCLUSIONS

A novel quantization-based data hiding method is proposed. The simulation results show that TCQ-IS method has better performance with respect to other well-known quantization based data hiding methods for certain input sources and channel noises, especially when distortion noise is high. Furthermore, embedding data using the magnitude of Discrete Fourier Transform coefficients of natural images by TCQ-IS is also implemented and it is observed that dense TCQ structures yield better results. As a future research optimal structures for certain attacks should be investigated.

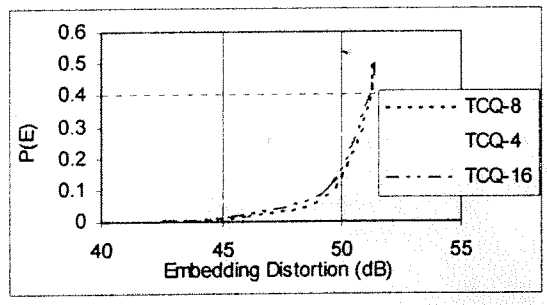


Fig.10. JPEG-80 Compression Attack Performance.

7. REFERENCES

[1] P. Moulin, "The role of information theory in watermarking and its application to image watermarking," invited paper, *Signal Processing.*, vol. 81, no. 6, pp. 1121-1139, June 2001.

[2] M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol.29, pp.439-441, May 1983.

[3] B. Chen, *Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems*, Phd. Dissertation, MIT, 2000.

[4] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no.4, May 2001.

[5] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between data hiding and distributed source coding," *Invited Paper, Proc. 33rd Annual Asilomar conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1999.

[6] J. Chou, S. S. Pradhan, L. El Ghaoui, and K. Ramchandran, "A robust optimization solution to the data hiding problem using distributed source coding principles," *Proc. SPIE conference, San Jose, CA, Jan. 2000.*

[7] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. Inform. Theory*, vol. 49, no.5, May2003.

[8] M.W. Marcellin and T.R. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Transactions on Communications*, vol. 38, pp. 82-93, January 1990.

[9] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol.28, pp.55-67, Jan. 1982.

EUROCON 2003 Ljubljana, Slovenia

Trellis Coded Quantization for Data Hiding

E. Esen[1, 2], A. A. Alatan[1, 2], and M. Aşkar[2]
[1]TÜBİTAK BİLTEN and [2]Department of Electrical-Electronics Eng., M.E.T.U.
Balgat 06531 Ankara, TURKEY
email: ersin.esen@bilten.metu.edu.tr

[1]*Abstract*— **The intrusion of information theoretic tools into the data hiding realm lead to the design and analysis of new blind detection methods. Although an extended analysis has already been built on different quantization-based data hiding methods, we propose another quantization-based method, which uses trellis coded quantization. The performance of the proposed method is compared against other well-known methods by simulations. The promising results show that the proposed method can be preferred in certain applications.**

Index Terms— **Data hiding, digital watermarking, fingerprinting, trellis coded quantization, TCQ, quantization index modulation, QIM.**

I. INTRODUCTION

DATA hiding found new application areas in new digital world, as significant need for digital watermarking and fingerprinting has emerged. Although the general framework of data hiding is much broader than the process of embedding a related message into the content itself, digital watermarking constitutes one of the most active research areas in image processing field, starting from the early 90's [1] with many proposed methods for this purpose [2].

In the early stages, mostly *informed detection (private watermarking)* has been worked on, compared to *blind detection (public watermarking)*. In the informed detection, the original content is used to decode the watermark, whereas this content is not available in blind detection. Although, the former problem intuitively appears to be simpler than the latter, some recent developments based on information theoretic tools [3], [5], caused blind detection problem to attract more attention. Costa's paper [5] is one of the most remarkable works, which results in the analysis and design of novel blind detection methods. Costa shows the equality between the capacity of a communications system with side information available to only encoder and that of the system with side information available to both encoder and decoder [5]. Considering watermarking as a communications problem with side information as shown Figure 1, where w is the data or the watermark to be embedded, s is the side information, x is the signal containing the hidden data or watermark, y is the signal available at the decoder after the channel utilization, $\hat{w}$ is the decoded data or watermark, and finally, applying the result of Costa, the capacity of the system is given as [5],

$$C = \frac{1}{2}\log(1+\frac{P_x}{N}) \ , \qquad (1)$$

where P_x is the transmitter power constraint and N is the channel noise (attack) power. As apparent in (1), the capacity does not depend on the side information, which is the signal in which data or watermark is embedded. Hence, it is possible to devise public watermarking systems achieving the same capacity with private systems.

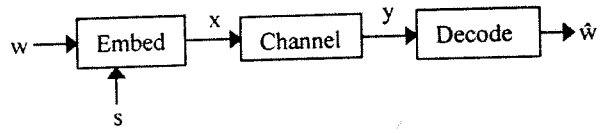


Fig. 1. Watermarking as a communications problem.

Inspired by the information theoretic tools, many public watermarking algorithms are developed and analyzed [6], [7], [8], [9]. Chen et.al. [6], [7] has introduced Quantization Index Modulation (QIM), as a generic class of data hiding methods. QIM embeds data into cover content using quantization. Different quantizers are used for different watermarks or data; i.e. quantizers are modulated according to the data. At the decoder, the hidden data is decoded using the knowledge of the QIM structure employed at the encoder; hence, the original content is not required. Chen et al [7] also showed that QIM structures are optimal under certain conditions. Another public data hiding method based on quantization is proposed in [8], [9]. Chou et.al. [10] use quantization after trellis coded modulation as a consequence of the observation of the duality between source coding and channel coding with side information.

In this paper, we propose yet another quantization-based public data hiding method. The method uses Trellis Coded Quantization (TCQ), which is described briefly in Section II. The method is presented in Section III. The robustness of the proposed methods is compared with other well-known methods through various simulations, which are given in Section IV. Section V concludes with some remarks.

II. TRELLIS CODED QUANTIZATION

TCQ [11] can be considered as a special case of trellis coding. The main ideas behind TCQ are due to trellis coded

This research is partially supported by Scientific and Technical Research Council of Turkey (TÜBİTAK) [Project # EEEAG 101E007]

modulation (TCM) [12]. TCQ employs a set of trellises and set partitioning ideas of TCM in order to achieve better distortion performance with low complexity.

TCQ uses a trellis and an associated codebook. A sample trellis of 4 states is shown in Figure 2, in accordance with its finite state machine and the codebook. In Figure 2, u is the input bit of the finite state machine, s_1, s_0 are the states, and o_1, o_2 are the outputs of this machine. The smooth lines in the trellis correspond to $u=0$ and the dashed lines correspond to $u=1$. Each branch of the trellis is associated with a subset D_i of the codebook, whose index i is determined by the output of the finite state machine. In order to design the codebook of TCQ, a scalar codebook C of size $2^{R+\dot{R}}$ is taken. R is the encoding rate and $\dot{R}$ represents the number of bits that specify the particular codeword in the selected subset. For the sample system in Figure 2, R is two bits per sample (bps) and $\dot{R}$ is equal to one. For an encoding rate of 2 bps, C is twice larger than the corresponding scalar quantizer. After C is determined, it is partitioned into $2^{\dot{R}+1}$ subsets, each of which has 2^{R-1} codewords.

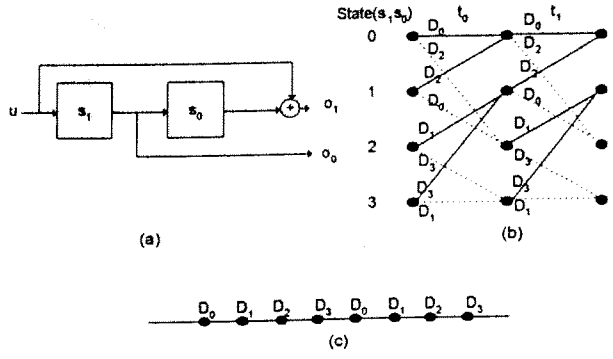


Fig. 2. TCQ structure

In order to quantize an input sequence s of length m, a trellis of m stages is used. Viterbi algorithm [16] is used to find the closest path to s among all possible trellis paths. $m(R-\dot{R})$ bits specify the trellis path and $m\dot{R}$ bits specify the codeword in the selected subset. Initial state is generally taken as 0, since for long data sequences ($m>>log_2N$, where N is the number of states) the effect of the initial state on Mean Square Error (MSE) is negligible [13].

Apart from the described fixed rate TCQ, other versions of TCQ can also be found [14], [15].

III. TCQ-IS

The proposed method makes use of TCQ for hiding data. The main observation behind the proposed method is the effect of the initial state on MSE. One can embed some information into the content by enforcing the selection of the initial state, accordingly. Since the data is hidden into the initial state (IS), the proposed method is denoted as Trellis Coded Quantization-Initial State (TCQ-IS).

After enforcing the selection of the initial state according to the data to be hidden, the closest trellis path is found using Viterbi algorithm, as in conventional TCQ.

Figure 3 shows a sample embedding process for TCQ-IS. In this illustrated example, $N=4$, $m=3$, $R=2$, $\dot{R}=1$. Since N is 4 one can hide 2 bits in the m input samples. Assuming, the data vector is selected as w={0,1}, the second state is chosen as the initial state. The corresponding closest path found by Viterbi algorithm is shown with thicker branches.

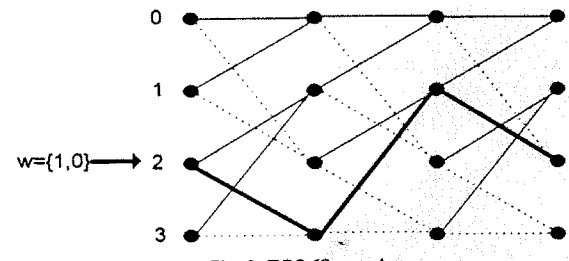
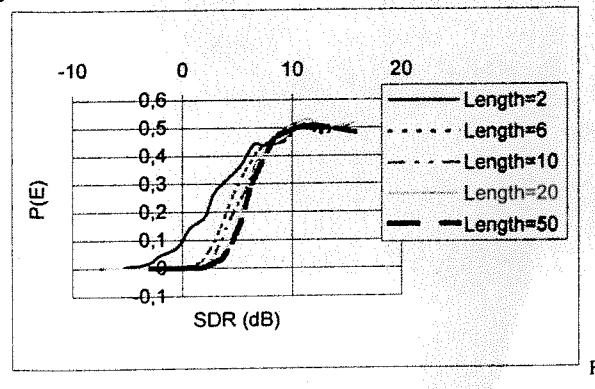


Fig. 3. TCQ-IS sample

The total number of bits that can be hidden by TCQ-IS is log_2N, which depends only on the number of states in the trellis. Embedding rate of TCQ-IS is log_2N/m bps. For a given m, one should use trellises with more states in order to hide more data.

In the decoding stage, all of the N states are considered as the candidate initial states. For each candidate initial state Viterbi algorithm is executed for the received input vector y. The computed MSE values are stored and the initial state with minimum MSE gives the desired embedded data, as the index of the corresponding state.

The robustness performance of the input data sequence against its length is shown in Figure 4. The robustness is measured by the relation between the probabilities of erroneous decoding versus the embedding distortion (signal to distortion ratio, SDR). Figure 4 displays the robustness of TCQ-IS with different lengths for Gaussian input source with zero mean, unit variance against a Gaussian channel noise of zero mean, 0.5 variance. As apparent from Figure 4, increasing data length improves the performance. However, beyond a limit the difference becomes indistinguishable.



F

ig. 4. Robustness of TCQ-IS for different lengths.

IV. SIMULATIONS

The robustness of TCQ-IS, for different input data sources and channel noises, is compared with two other quantization-based data hiding methods, namely Dither Modulation (DM) [6] and TCQ Path Selection (TCQ-PS) [8], [9].

The performances of all these methods are compared by using the relation between the embedding distortion (SDR), which is caused during the data embedding process, and the probability of error, P(E), at the decoder. The embedding distortion is measured as the ratio of the input signal power to the embedding distortion power in logarithmic scale. The probability of error is computed as the ratio of erroneously decoded data bits to the number of total data bits. The final values are computed by averaging the results for 500 random experiments. Two input sources are used to embed data: uniform in the range (-1,1) and Gaussian with zero mean and unit variance. The input source powers are kept constant while the channel noise power is varied in order to see the performance of the methods with respect to different channel characteristics.

The TCQ structure in Figure 2 is utilized for TCQ-IS and TCQ-PS during all of the simulations. The codebook is chosen so that it covers all input data range for uniform source. For Gaussian source, the selected range contains most of the signal energy.

Due to the requirement for embedding rate equality, the simulations are performed in two parts. First, TCQ-IS and DM are compared with each other. Then, all of the three methods are compared with proper input data sequence length adjustment. Before proceeding to simulations the methods used for comparison are described briefly in the following subsections.

A. *Dither Modulation (DM)*

DM is a member of the QIM family. DM employs a quantizer set in which the quantization cells and the reconstruction points are the shifted versions of any other quantizer in the set [6]. The dither vector d is modulated with the watermark w. Then, using the corresponding dither vector for a given w, the embedding function is,

$$x(s,w) = q(x + d(w)) - d(w)\,, \qquad (2)$$

where q is the selected base quantizer. Figure 5 shows a sample DM structure. A uniform quantizer with step size Δ is used as the base quantizer. The dither vectors are $-\Delta/4$ and $+\Delta/4$ for $w=0$ and $w=1$, respectively. The hidden data is decoded by computing the distance between the reconstruction levels of each quantizer in the set. The index of the quantizer with the minimum distance gives the decoded data. According to the desired embedding rate, input data s can be used individually or sequentially in conjunction with majority voting in the decoding stage.

B. *TCQ-PS*

TCQ-PS is based on the TCM-TCQ scheme described in [8], [9]. Using the same trellis structure in TCQ-IS, the coset selection in which the input signal s is to be quantized [8], [9] reduces to the trellis path determination according to the data w. Once the path is determined s is quantized using the corresponding D_i. In that sense, TCQ-PS is also a member of

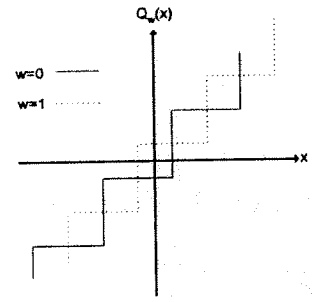


Fig. 5. A sample Dither Modulation structure.

QIM family. However, in this case, the quantizers are selected according to the trellis. Figure 5 shows a sample for TCQ-PS. The trellis path is determined with respect to the data w={1,0,1} starting from initial state 0. The data is quantized using the associated subsets of the branches in the order dictated by the trellis. In this scheme, embedding rate is 1 bps. At the decoding side, Viterbi algorithm is used to quantize the received signal y using the given trellis structure. Once the best path is found, the data is decoded starting from the initial state 0.

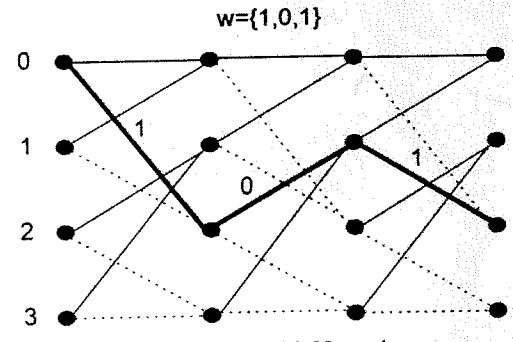


Fig. 6. A TCQ-PS sample.

C. *TCQ-IS vs. DM*

a. *Uniform Source*

In these simulations, input data length is selected as 10. Since the structure for TCQ-IS dictates an embedding rate of 2/10, 1 bit is hidden into every 5 input samples for DM. Majority voting is used during decoding. Figures 7, 8, and 9 shows the performance of the methods for Gaussian channel noises with standard deviations 0.5, 1, and 2, respectively. For all cases TCQ-IS has a better performance, that is, for a given embedding distortion the probability of error of TCQ-IS has smaller values or for a given probability of error, the required distortion is smaller. Furthermore, the plots also indicate that the performance for TCQ-IS improves, as the channel noise power increases.

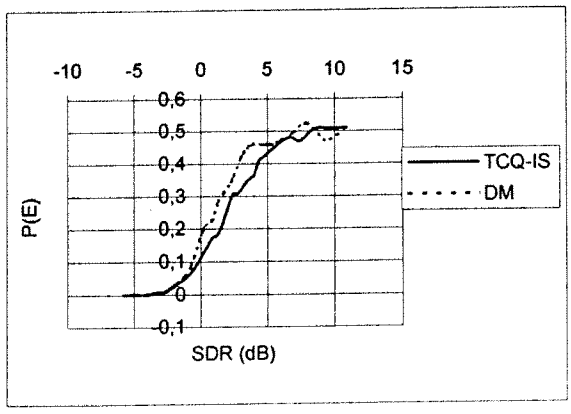


Fig. 7. TCQ-IS vs. DM for uniform source: channel noise G(0,0.5)

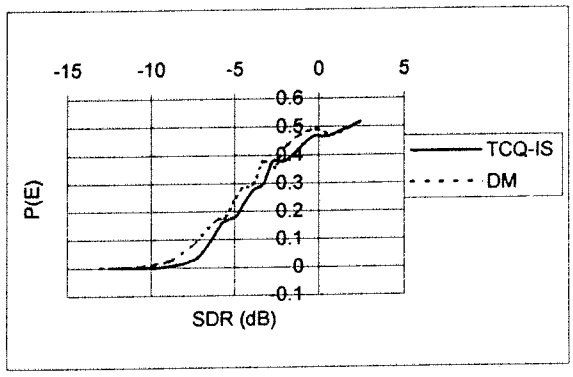


Fig. 8. TCQ-IS vs. DM for uniform source: channel noise G(0,1)

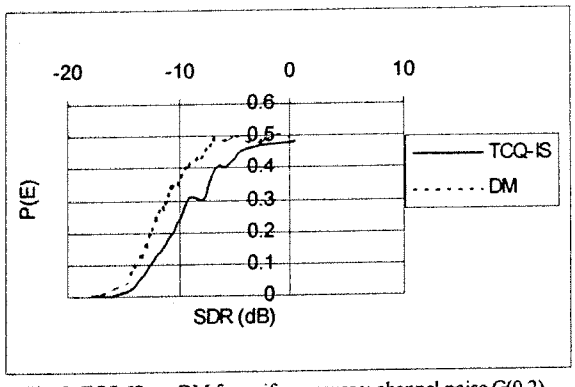


Fig. 9. TCQ-IS vs. DM for uniform source: channel noise G(0,2)

b. Gaussian Source

The same experiments in (a) are repeated for Gaussian inputs. Figures 10, 11, and 12 show the same results for Gaussian channel noises with standard deviations 0.5, 1, and 2, respectively. For low channel noise powers, the two methods have similar performances. As the channel noise power increases, the performance of TCQ-IS gets better.

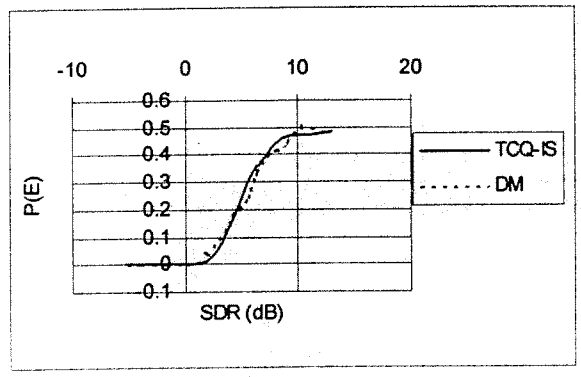


Fig. 10. TCQ-IS vs. DM for Gaussian source: channel noise G(0,0.5)

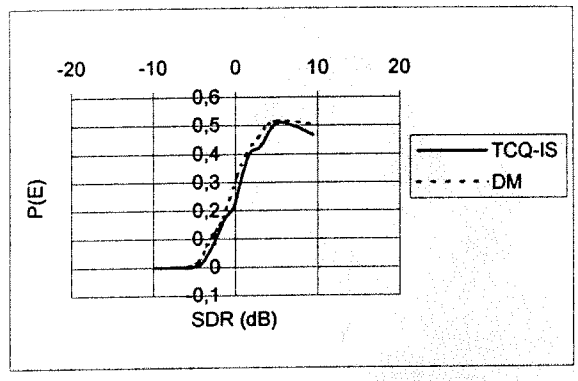


Fig. 11. TCQ-IS vs. DM for Gaussian source: channel noise G(0,1)

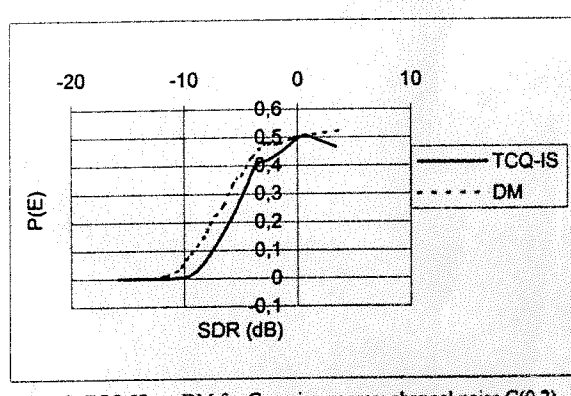


Fig.12. TCQ-IS vs. DM for Gaussian source: channel noise G(0,2)

D. Overall Comparison for the methods

The same experiments in (C) are conducted with the exception that, TCQ-IS operates on input samples in lengths of 2, in order to attain the same embedding rate 10/10 bps for TCQ-PS. In this case, DM uses 1 sample to hide 1 bit.

a. Uniform Source

The results for Gaussian channel noises with standard deviations 0.5 and 1 are shown in Figures 13 and 14, respectively. For low noise case, TCQ-IS and DM have similar performances, which are still better than TCQ-PS. For high noise case, the performance of TCQ-IS is superior to all other methods.

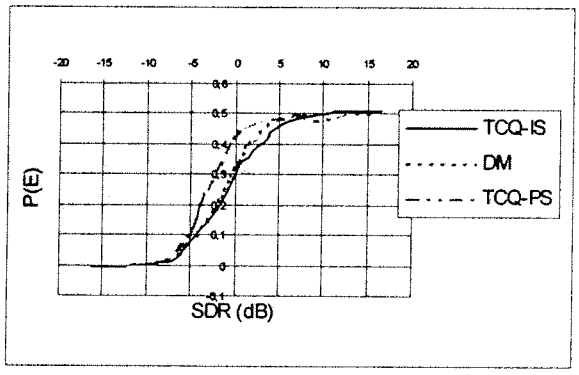


Fig. 13. Uniform source under channel noise G(0,0.5)

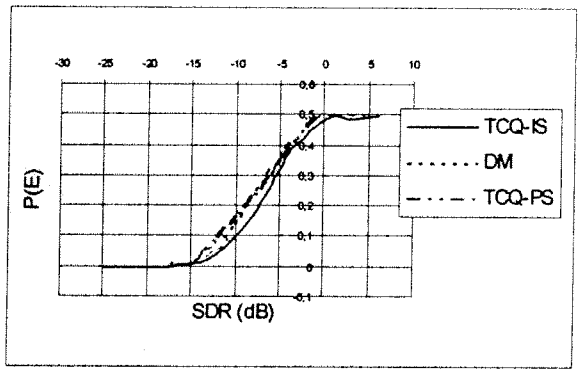


Fig.14. Uniform source under hannel noise G(0,1)

b. Gaussian Source

Figures 15 and 16 show the results for Gaussian channel noises with standard deviations 0.5 and 1. TCQ-IS and DM have similar performances in both cases. For high noise case, the superiority of TCQ-IS and DM over TCQ-PS becomes indistinguishable.

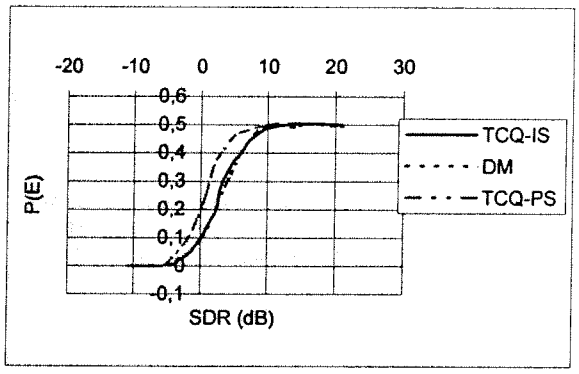


Fig. 15. Gaussian source under channel noise G(0,0.5)

V. CONCLUSION

The simulation results show that proposed TCQ-IS method has superior performance with respect to other well-known quantization based data hiding methods for certain input sources and channel noises, especially when the distortion noise is high. Hence, one can state that TCQ-IS can be used in the applications, where such conditions are met. Embedding data into the magnitude of discrete Fourier coefficients of natural images can be a candidate typical approach. The effect of using higher number of states and codebook design, which makes use of the source entropy, should be investigated in such an approach, as a further research.

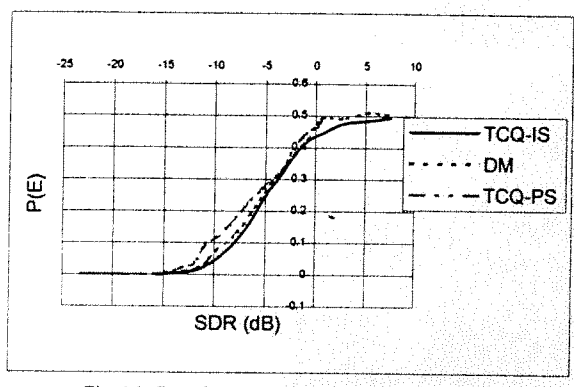


Fig. 16. Gaussian source under channel noise G(0,1)

REFERENCES

[1] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding- a survey," *Proceedings of the IEEE*, vol. 87, no. 7, pp.1062-1078, July 1999.

[2] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*, Academic Press:USA, 2002.

[3] P. Moulin, "The role of information theory in watermarking and its application to image watermarking," invited paper, *Signal Processing*, vol. 81, no. 6, pp. 1121-1139, June 2001.

[4] A. Sequeira, D. Kundur, "Communication and Information Theory in Watermarking: A Survey," *Multimedia Systems and Applications IV*, A. G. Tescher, B. Vasudev, and V. M. Bove, eds., Proc. SPIE (vol. 4518), pp. 216-227, Denver, Colorado, August 2001.

[5] M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol.29, pp.439-441, May 1983.

[6] B. Chen, *Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems*, Phd. Dissertation, MIT, 2000.

[7] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no.4, May 2001.

[8] J. Chou, S. Sandeep Pradhan, and K. Ramchandran, "On the duality between data hiding and distributed source coding," *Invited Paper, Proc. 33rd Annual Asilomar conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1999.

[9] J. Chou, S. Sandeep Pradhan, L. El Ghaoui, and K. Ramchandran, "A robust optimization solution to the data hiding problem using distributed source coding principles," *Proc. SPIE conference, San Jose, CA, Jan. 2000.*

[10] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. Inform. Theory*, vol. 49, no.5, May2003.

[11] M.W. Marcellin and T.R. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Transactions on Communications*, vol. 38, pp. 82–93, January 1990.

[12] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol.28, pp.55-67, Jan. 1982.

[13] D.S. Taubman and M.W. Marcellin, *JPEG2000: Image compression fundamentals, standards and practice*, Kluwer Academic Pub., 2002.

[14] J.H. Kasner, M.W. Marcellin, and B.R. Hunt, "Universal trellis coded quantization," *IEEE Transactions on Image Processing*, Vol. 8, No. 12, pp. 1677–1687, December 1999.

[15] H. Brunk and N. Farvardin, "Embedded trellis coded quantization," *Data Compression Conference*, pp. 93-102, 1998.

[16] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp.268-278, Mar. 1973.

ERROR CONCEALMENT OF VIDEO SEQUENCES BY DATA HIDING

Ayhan Yılmaz and A. Aydın Alatan
Electrical and Electronics Engineering Department
M.E.T.U., Balgat, 06531 Ankara, TURKEY
E-mail: {ayhan, alatan}@eee.metu.edu.tr

ABSTRACT

A complete error resilient video transmission codec is presented, utilizing imperceptible embedded information for combined detecting, resynchronization and reconstruction of the errors and lost data. Utilization of data hiding for this problem provides a reserve information on video at the receiver while unchanging the transmitted bit-stream syntax; hence, improves the reconstruction video quality without significant extra channel utilization. A spatial domain error recovery technique, which hides edge orientation information of a block, and a resynchronization technique, which embeds bit-length of a block into other blocks are combined, as well as some parity information about the hidden data, to conceal channel errors on intra-coded frames of a video sequence. The inter-coded frames are recovered by hiding motion vector information into the next frames. The simulation results show that the proposed approach performs superior to conventional approaches for concealing the errors in binary symmetric channels, especially for higher bit-rates and error-rates.

1. INTRODUCTION

Transmission of video signals over noisy wireless channels may cause inevitable errors that might severely degrade the visual message. Error concealment techniques can be classified into 3 major groups with the following properties [1]: an interaction between the encoder and decoder, as a re-send signal, or post-processing operations at the decoder to recover lost information, or leaving some extra redundancy at the encoder to minimize the reconstruction error.

All these approaches can be combined by hiding some imperceptible information to be useful during error concealment. In this way, hidden information is not only transmitted through a (secret) channel from encoder to decoder, "sending back" some lost information, but also alleviates some burden on post-processing. Moreover, the extra hidden information and its small visual loss might be equivalent decreasing the source bit-rate for obtaining the same visual quality and utilizing error control codes as a result of the bit savings at the encoder.

This radical approach in video error concealment is a result of steganography, a new technique for making imperceptible modifications on the media, mostly utilized for copyright protection and other security-based applications [2,3]. Recently, erro[1]r resilient video transmission has become a new application area for data hiding, as some novel concealment methods are proposed [4-10]. These methods are examined in the next section.

It should be emphasized that the hidden information can be transmitted with a very small bit-rate overhead in the bit-stream. The standard receivers unaware of such hidden information will be unaffected and decode the bit-stream, successfully (i.e. backward compatibility between the bit-streams and conventional decoders). Obviously, the price, one pays for this additional gain, is an increasing complexity at the decoders and a small loss in visual quality.

The main motivation of this research is to demonstrate the advantages of data hiding over conventional methods in error concealment problem for video transmission. None of the previous approaches has proposed a complete video codec with detection, synchronization and recovery (reconstruction) capabilities together based on data hiding and tested under noisy channel conditions. In our simulation results, the performance comparisons are given for a number of sequences under various channel conditions at different bit-rates.

2. ERROR CONCEALMENT USING DATA HIDING

All state-of-the-art standard-based video codecs are block-based, consisting of intra- and inter-coded frames. In these systems, the conventional approaches conceal errors by either reconstructing the lost block with the smoothness property of the intensities in an intra-coded frame, or estimating a lost motion vector from block motion vectors of its spatio-temporal neighbors to compensate for the lost block in an inter-coded frame [1]. Except for the damage of any header information in the bit-stream, the bit-errors usually destroy the data only in a single block or sometimes in all the blocks within the rest of the row of macroblocks (slice). In case of such synchronization loses, the reversible variable-length coding can be the only solution with limited capabilities in some recent standards (e.g. MPEG-4).

The error concealment methods utilizing data hiding technology can be broadly classified into 3 approaches: detection of errors [4,5], resynchronization after detected errors [6] and reconstruction of the intensities for the lost blocks [7-10].

Error detection schemes based on data hiding either hide the parity check codes of the macroblocks [4] or modify DCT coefficients according to their frequency location [5]. In [4], the LSB of the sum all non-zero DCT coefficients is hidden to the next frame as a parity to detect errors. As a different approach, the value of a DCT coefficient is forced to become even (or odd), according to its DCT frequency, indexed in zigzag order [5].

This research is supported by Scientific and Technical Research Council of Turkey (TÜBİTAK) [Project # EEEAG 101E007]

The synchronization loss within a slice can be recovered by hiding the bit-length value of each macroblock (in binary form) into another slice [6]. Hence, if the video codec is able to detect an error in a slice, the recovered macroblock lengths are utilized to decode the bit-stream block-by-block in a reverse order from the end of slice. In this approach [6], the detection performance is limited with video codec capabilities.

Finally, methods for the reconstruction of intensities should be examined for intra- and inter-coded frames, separately. In case of still images (intra-coded frames), finding a "good summary" for the intensity information of the lost block is the critical point. Block intensity information is approximated by using either edge-direction information [7] or a low-quality coded (high compression) version [8]. It is shown that interpolation of lost intensities along a major edge gives much superior results against well-known bilinear interpolation approach at the decoder [11]. Instead of trying to find the edge direction of a lost block from its neighbors in a suboptimal manner at the decoder, hiding the quantized direction information into another block at the encoder is preferred [7].

For inter-frame error concealment, motion vector of a block is obviously the most valuable information. Instead of trying to estimate the lost motion vectors (MVs) by replacing them with the average/median of the MVs from spatial/temporal adjacent blocks, as in conventional approaches [1], motion vector information is simply hidden into other blocks/frames [9,10]. In order to minimize the number of bits to hide, the `modula-2` addition of the motion vector bit-streams for all slices in one frame [9] or consecutive frames [10] is hidden into motion vectors [9] or DCT coefficients [10] of the next frame.

All the attractive properties of error concealment using data hiding can be merged to obtain a system that is capable of jointly detecting, resynchronizing and recovering errors in video sequences. A novel method to enhance error resilience properties of a ITU H.263+ codec using hidden information is explained in the next section.

3. PROPOSED METHOD

In the proposed method, the data is always embedded into the frames by *even-odd signaling* [2] of the DCT coefficients. This approach gives maximum capacity with minimum robustness, which is acceptable for our application.

In order to achieve successful error concealment, the exact location of the error, i.e. damaged block, should be detected as a first step. After detecting the damaged block, synchronization must be established back in order to prevent the propagation of the error to the other blocks. The final step is the reconstruction of the intensities for the damaged block to finalize error concealment. Therefore, the three main issues for successful error concealment are error detection, resynchronization and reconstruction (recovery) of the damaged block.

3.1. Intra-coded frame Concealment

Following the previous approaches [4-10], both edge- direction information [7] and macroblock bit-length values [6] are necessary to solve all three problems. While bit-length value (8-13 bits, determined according to the bit-rate) is hidden to a previous block for resynchronization, quantized edge-direction information (4 bits) is embedded into the upper block to be used in case of damage (Fig.1). Prior to embedding edge orientation, the block is tested for being classified as an *edge block* by computing the gradient magnitude and this single bit is also stored to select bilinear interpolation for smooth block recovery. During decoding, the system is not allowed to decode bits more than the number that is dictated by the hidden value in the previous block. Finally, edge-direction and bit-length values are crosschecked to detect bit-errors (Fig.1).

3.1.1. Errors damaging the hidden data with no visual effect

Although, the hidden bits are enough to detect an error as an inconsistency between a block and its hidden data, they do not determine the errors, which do not change the visual information (i.e. edge direction of the block). Such errors are very likely to corrupt the hidden information in the block. This problem for error detection by using hidden information is mostly neglected in the previous methods. Hence, in case of an inconsistency between block and the hidden data, a single-bit parity of the macroblock bit-stream is embedded to verify the reliability of the hidden data (Fig.1).

3.1.2. Overconcealment

After successful error detection, another important problem is to "measure" the visual damage at a block before recovery, since it is possible to have a very small visual error in the block, undetected by the codec itself, but "successfully" detected by the proposed system. In such a case, the edge-direction based recovery technique tries to reconstruct the block, which has originally negligible visual degradation, while discarding all the available information. Obviously, the reconstruction quality usually turns out to be inferior compared to the erroneous block.

We denote this situation as *overconcealment* and it is avoided by hiding `modula-2` sum of 2-bit MSBs of the current block coefficients, as a visual loss parity. It is assumed that in case of visually unacceptable errors, 2-bit MSBs are changed and this change can be detected by 2-bit parity information hidden into the previous block (Fig.1). Note that the whole idea is not to conceal, if there is not sufficient visual loss.

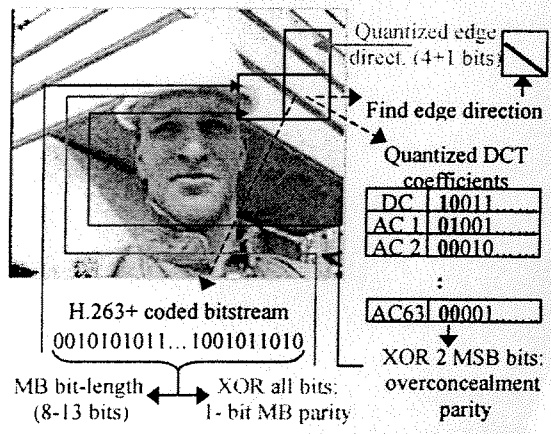


Figure 1. Hiding data for error concealment for intra-frames

3.2. Inter-coded frame Concealment

The motion vector of each block is strictly necessary during inter-frame recovery and obviously, this data should be hidden into other blocks.

Following [9,10], in the proposed approach, bit-stream of the differential Huffman coded MVs of each row, as well as their bit-length, are embedded into the motion compensated residual DCT coefficients of the corresponding row in the next frame (see Fig.2). Obviously, if there are errors in the same rows of the successive frames, then the hidden information is not useful anymore.

3.2.1. Measuring visual loss and Locating error

Compared to intra-coded frame case, similar problems still exist for the inter-coded frames. In this case, a problem analogous to overconcealment also exists and one should decide whether to add motion compensated residual error on motion-based prediction. If a visually significant error is detected in the motion compensated error bit-stream, then the decoded coefficients are not used to reconstruct the current block. Visual significance of the error is again tested by 2-bit MSBs of the DCT coefficients.

On the other hand, in order to find the location of error, all the hidden data is protected by a checksum (5 bit). This gives the opportunity to comprehend the reliability of hidden data (Fig.2).

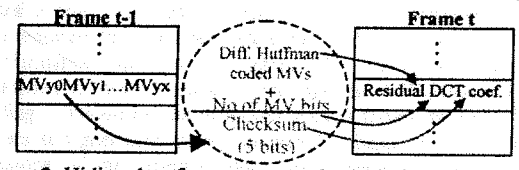


Figure 2. Hiding data for error concealment in inter-frames

3.3. The Proposed Algorithm

An overview block diagram of the algorithm is given in Fig.3 (a) and (b) for intra- and inter-coded frames, respectively. In both versions, there are consecutive error detection stages. The internal error detection mechanism of H.263+ is used as default in both inter-and intra-coded versions. For the errors invisible to codec, the major detection test for intra-coded frames is synchronization and parity check, whereas the inter-coded version controls the checksum information. In addition, the intra-coded case checks overconcealment before any reconstruction.

Apart from this overview in Fig. 3, there are also some implementation details about the proposed algorithm, such as checking continuously the reliability of the hidden data or using edge information to check errors, if the hidden data for synchronization is not available.

4. SIMULATIONS

During experiments, a Binary Symmetric Channel (BSC) is simulated for different Bit Error Rates (BER). By using a fully implemented encoder-decoder pair, input data is first compressed with a baseline H.263+ encoder and during this time, the hidden info is also embedded into the bit-stream. Then, the bit-stream is passed through a BSC, and finally, the erroneous bit-stream is decoded using a modified H.263+ decoder, capable of error concealment from hidden data. During simulations, *Foreman* and *Carphone* (400 frames and QCIF size) are utilized and the bit-streams are test with BSC for different error patterns 100 times. The visual reconstruction quality is determined in terms of Peak Signal-to-Noise ratio (PSNR).

The results of simulations are summarized in Tables 1 and 2. In these tables, PSNR after compression, data hiding and

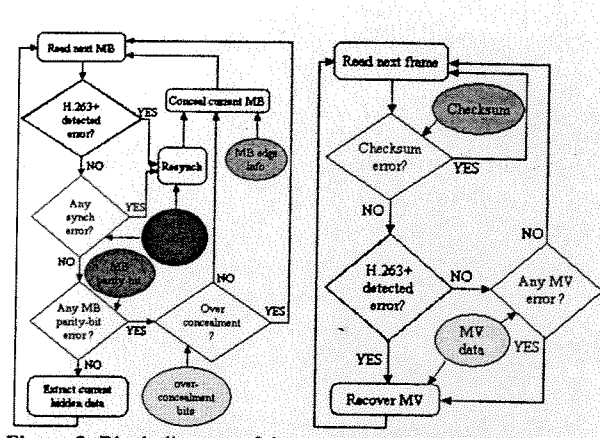


Figure 3: Block diagram of the proposed algorithm for (a) intra- and (b) inter-coded frames (shaded ellipses show hidden data).

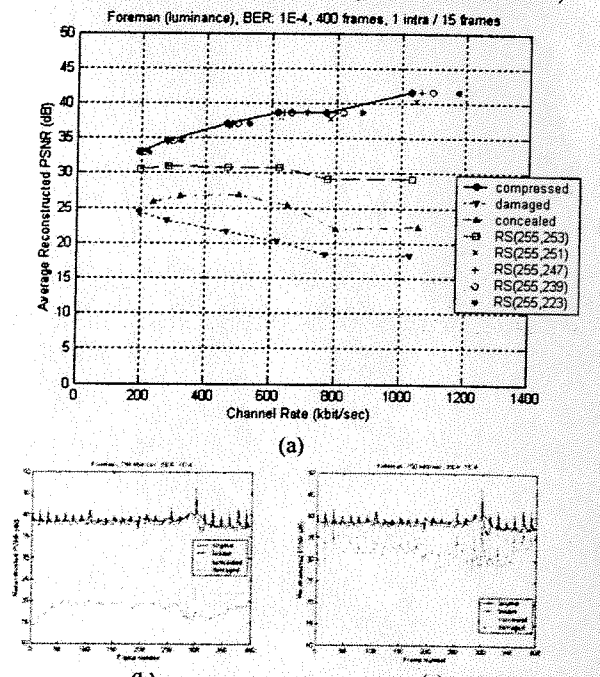


Figure 4: *Foreman* (a) PSNR vs. bit-rate for BER: 10^{-4}, (b) PSNR vs. frames for bit-rate=750 Kb/s, BER: 10^{-4}, (c) BER=10^{-5}

resulting average PSNRs for damaged and concealed sequences are presented for three different bit-rates. Note that the damaged sequences are obtained using the error resilience properties of H.263+ baseline codec, in which the error reconstruction is simply achieved by replacing a lost block from the same location in the previous frame.

The system is also compared with (255,253), (255, 251), (255,247), (255, 239), and (255, 223) Reed-Solomon (RS) codes for 6 different bit-rates. The results are shown in Fig.4 (a) for Foreman sequence and BER 10^{-4}.

The PSNR plot is shown in Fig.4 (b) and (c) for Foreman sequence for BER 10^{-5}, 10^{-4} and bit-rate 750Kb/s. From these

tables and Fig.4, it is clear that the performance of the proposed algorithm improves for higher error rates (BER=10^{-4}) and an average improvement of about 4dB is achieved for this BER over the baseline codec. However, the well-known RS codes give better results compared to the proposed method. For the simulated bit-rates, data hiding causes about 1dB visual loss. On the other hand, the performance of the proposed algorithm gets better, as the encoding bit-rates increases. For BER=10^{-5}, in low bit-rates, there is an over protection by data hiding, resulting with a significant decrease in PSNR after data hiding.

5. CONCLUSIONS

A novel video error concealment method is proposed, jointly achieving detection, resynchronization and recovery using data hiding. The system combines different types of hidden information in order to obtain better reconstruction quality. Simulations on video sequences indicate a significant improvement against conventional techniques.

The reason of observing better performance at higher bit-rates is due to enough number of non-zero coefficients to hide the required data for inter-coded frames. In order to improve the performance for lower bit-rates, the algorithm should be modified to use bit-planes other than the LSB-plane for data hiding. Obviously, the proposed system shows its resilience to errors at higher error-rates, compared to baseline codec. Robust versions of the original codec are expected to decrease the performance gap with the proposed method.

It should be noted that not all the blocks have the same characteristics from reconstruction point of view. The simulations show that blocks without a major single edge (such as highly textured areas) cannot be interpolated successfully via edge-based interpolation.

The major drawback for error concealment using data hiding is due to the fundamental dilemma for finding the source of the error, in case of a conflict between hidden and decoded data. Using a number of extra parity bits partially solve these problems, since these solutions come along with the assumptions on the locations of bit-errors (e.g. bit-errors should be as far as possible). Utilization of more sophisticated error detection mechanisms (e.g. CRC), instead of simple parity bits may solve this problem, while increasing the number of bits to hide.

REFERENCES

[1] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: A review," Proc. of the IEEE, vol. 86, no. 5, pp. 974-997, May 1998.

[2] M. Wu and B. Liu, "Watermarking for image authentication," ICIP'98, vol. 2, pp. 437-441.

[3] M. Ramkumar, A. N. Akansu, and A. A. Alatan, "On the choice of transforms for data hiding in compressed video," Proc. of IEEE ICASSP '99, pp. 3049-3052, 1999.

[4] T. S. Wang, P.-C. Chang, C.-W. Tang, H.-M. Hang, and T. Chiang, "An error detection scheme using data embedding for H.263 compatible video coding," ISO MPEG, N6340, July 2000.

[5] A. Piva, R. Caldelli, V. Cappellini, A. De Rosa, "Data hiding for transmission error detection in H.263 video," Tyrrhenian Int Workshop on Digital Communications, IWDC 2002, Capri, Italy, September 8-11, 2002.

[6] D. L. Robie and R. M. Mersereau, "Video error correction using steganography" Proc. of ICIP, pp.930-933, October 2001.

[7] P. Yin, B. Liu, and H. H. Yu, "Error concealment using data hiding," Proc. of. ICASSP, vol. 3, pp. 1453-1456, May 2001.

[8] S. Yafei, Z. Li, W. Guowei, and L. Xinggang, "Reconstruction of missing blocks in image transmission by using self-embedding," Proc. of Int. Symp. on Intel. Multim., Video, and Speech Proc., p. 535-538, May 2001.

[9] J. Song and K. J. R. Liu, "A data embedding scheme for H.263 compatible video coding," IEEE Int Symp on Circuits and Systems, vol. 4, pp. 390-393, June 1999.

[10] P. Yin, M. Wu, and B. Liu, "A robust error resilient approach for MPEG video transmission over internet," Visual Communication and Image Processing, SPIE 2002, vol. 4671, pp. 103-111, January 2002.

[11] W. Zeng, and B. Liu, "Geometric-structure-based error concealment with novel applications in block-based low bit rate coding," IEEE Trans on Circuits and Systems for Video Tech, vol. 9, no. 4, pp. 648-665, June 1999.

Average PSNR (dB)		Bit rate: 1 Mbit/sec			Bit rate: 500 kbit/sec			Bit rate: 200 kbit/sec		
		Y	U	V	Y	U	V	Y	U	V
Compressed		41.56	43.57	44.96	37.09	40.38	41.73	33.02	37.99	38.91
Compressed + data hiding		40.80	43.11	44.24	36.00	39.96	41.24	29.59	37.35	37.77
BER: 1E-4	Damaged	18.17	21.88	21.56	21.50	25.24	24.96	24.38	28.87	28.58
	Concealed	22.30	29.81	29.67	26.86	34.87	35.49	25.74	33.00	33.23
BER: 1E-5	Damaged	32.62	37.34	38.01	33.31	37.33	38.32	31.66	36.71	37.47
	Concealed	35.80	40.49	41.34	34.76	39.59	40.81	29.07	36.54	36.92

Table 1. Average PSNR values for *Foreman* sequence for different bit-rates and BERs.

Average PSNR (dB)		Bit rate: 850 kbit/sec			Bit rate: 400 kbit/sec			Bit rate: 200 kbit/sec		
		Y	U	V	Y	U	V	Y	U	V
Compressed		42.69	44.54	45.32	38.38	41.53	42.32	34.86	39.40	40.12
Compressed + data hiding		41.65	44.02	44.83	36.39	40.85	41.61	31.38	38.67	39.13
BER: 1E-4	Damaged	19.50	22.55	22.45	22.56	25.66	25.69	25.10	29.22	29.33
	Concealed	24.10	31.02	30.89	27.69	35.79	35.58	26.56	33.23	32.95
BER: 1E-5	Damaged	35.25	38.82	39.24	34.72	38.01	38.96	33.45	38.04	38.66
	Concealed	36.43	40.63	40.98	35.27	40.40	41.23	30.69	37.75	38.00

Table 2. Average PSNR values for *Carphone* sequence for different bit-rates and BERs

COMBINED CONCEALMENT, SYNCHRONIZATION, AND ERROR DETECTION USING DATA HIDING

A. YILMAZ[1], E. ESEN[1,2] AND A. A. ALATAN[1,2]
[1]*Department of Electrical and Electronics Engineering, M.E.T.U.,*
[2]*TÜBİTAK BİLTEN, Balgat 06531 Ankara, Turkey*
e-mail: {ayhan, alatan}@eee.metu.edu.tr, ersin.esen@bilten.metu.edu.tr

Utilization of data hiding for concealment of video transmission errors provides a reserve information about the video to the receiver while unchanging the transmitted bit-stream syntax; hence, improves the reconstruction video quality without extra channel utilization. A spatial domain error concealment technique, which hides edge orientation information of a block, and a resynchronization technique, which embeds bit length of a block into other blocks are composed. The proposed method also uses these techniques for detecting errors. Simulation results show that these approaches cooperate quite well in concealing the errors.

1. Introduction

In order to handle the transmission errors of digital signals over noisy wireless channels at the receiver side, some error concealment techniques have been proposed [1]. These techniques try to recover the lost data by an interaction between the encoder and decoder, as a re-send signal [1], or post-processing operations to recover lost information at the decoder [5], or leaving some extra redundancy at the encoder to minimize the reconstruction error [1]. Instead of permitting some redundancy during source coding, one option is to hide some imperceptible information to be useful during error concealment.

Data hiding is a new generation technique of making imperceptible modifications on the media [3] and the hidden information can be transmitted without a bit-rate overhead in the bit-stream of the compression standard being used. Hiding information, not only works as a hidden channel from encoder to decoder, but also alleviates some burden on post-processing.

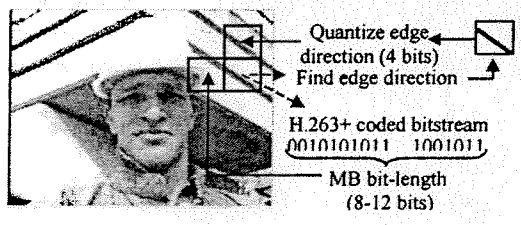
2. Related Work On Error Concealment By Data Hiding

Conventional error concealment methods are achieved in a post-processing stage and are mainly based on the smoothness property for the image frames. Recovery of the lost blocks is usually performed in spatial [5] or frequency domain [1]. However, if a lost block is not sufficiently smooth, then the interpolation of this block from its neighbors is a suboptimal solution. Data

hiding, which simply serves for transporting this lost information to the receiver, can be a good alternative solution with some bottlenecks.

Since interpolation along the edge direction gives superior results compared to conventional approaches, edge direction information is worth to hide [5]. In [2], the edge orientation of each block is embedded into a *companion* block (Fig. 1) by the help of any data hiding technique. Another problem arising in error concealment is due to the loss of synchronization in the bit stream at the decoder. One solution is to hide the bit length data of a block into neighbors [4].

Figure 1. Embedding edge orientation and bit length



3. Proposed Method

In order to achieve successful error concealment, the exact location of the error, i.e. damaged block, should be detected as a first step. After detecting the damaged block, synchronization must be established back in order to prevent the propagation of the error to the other blocks. The next step is the reconstruction of the intensities for the damaged block.

In the proposed method, edge direction information and bit length data are necessary to realize the error recovery in all 3 issues. Edge direction information is embedded into one upper block of the current block and bit length data is hidden into the previous block. For embedding the edge orientation, the block is first classified as an *edge block* by computing the gradient magnitude. The angles, whose gradient magnitudes are above a threshold, are quantized into 16 equally spaced directions (4 bits). Obviously, a single message bit should also be hidden to indicate the type of the block, i.e. an edge or a smooth block (Fig.1). For hiding the bit-length data, the number of bits used for the current block is determined during encoding and this value is embedded after a conversion into binary representation (Fig.1). The proposed method requires 8 to 12 bits according to the bit rate (i.e. quantization parameter). Data hiding is achieved by simple "even-odd" signaling [2,4].

3.1. *Error Detection*

Both edge direction information and bit length data is used for error detection. For each decoded block, its edge direction information is calculated once again

at the decoder and compared with the information hidden in the upper block. After the edge direction test, the bit length data is checked as second stage. After a block is decoded, total number of bits read from the bit stream and the value hidden in the previous block are compared.

3.2. *Resynchronization*

In order to resynchronize at the decoder, the bit length data is utilized again. The difference between the hidden and decoded bit numbers is calculated and the decoder skips that amount of bits in order to start decoding from a new undamaged block. In this way, without having macroblock headers, the system is able to synchronize itself at the start of each macroblock.

3.3. *Reconstruction*

For every block, edge direction information is extracted from the blocks in the upper slice. When an error is detected in a block, its edge direction information is checked whether it is an edge or a smooth block. If it is an edge block, then it is interpolated from two neighboring blocks along its edge direction. Otherwise, for a smooth block, simple bilinear interpolation technique is applied.

4. Simulations

ITU H.263+ codec is utilized for *Foreman* sequence during the experiments. After H.263+ compression, single bit errors are imposed on the resultant bit-stream. It should be noted that the current system considers only errors that affect the intra frames. A system, considers inter frames by hiding motion vector information, is still under development.

In Fig. 2(a), (b) and (c), the first frames of the original, compressed and data hidden videos are shown respectively. In Fig. 2(d), a typical result due to a random 1-bit error in the bit stream (intra-frame) is given. The concealment in Fig. 2(e) is made by bilinear interpolation only and H.263+ decoder itself detects the error. Fig. 2(f) is concealed by the proposed method. The decoder perfectly detects the error, synchronizes the next block and reconstructs the damaged block. Table 1 summarized the average PSNR values for the whole (22 times single-bit-error) experiments.

5. Conclusions

In this paper, a novel error concealment method using data hiding is proposed. The system combines two previous methods in this area in order to obtain much better reconstruction quality considering detection, resynchronization and

concealment together. However, all blocks do not have the same characteristics from concealment point of view. The simulations show that the blocks without a major single edge (such as highly textured areas) cannot be interpolated successfully via edge-interpolation.

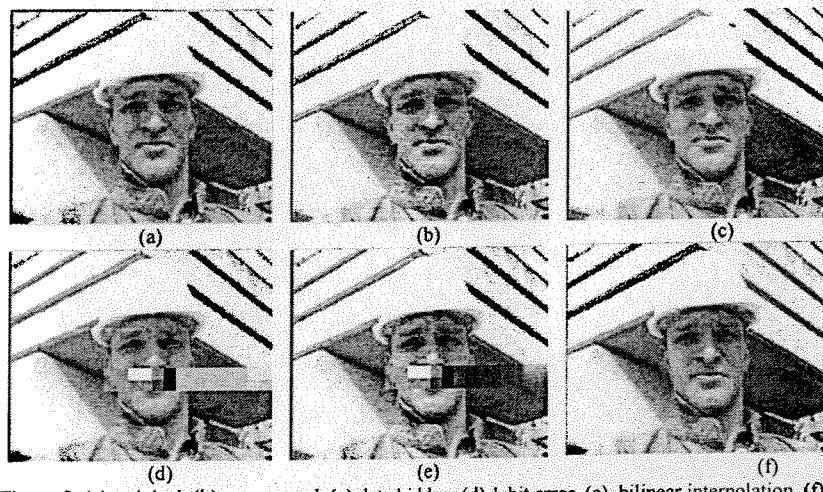


Figure 2. (a): original, (b) compressed, (c) data hidden, (d) 1-bit error, (e) bilinear interpolation, (f) proposed method.

Table 1. Average PSNR values.

Average PSNR (dB)	Y	U	V
After H.263+ compression (900 Kbit/s)	43.58	44.70	46.72
After data hiding	42.93	44.47	45.76
After transmission with errors	29.08	29.52	31.02
After concealment using bilinear interpolation	32.81	36.70	37.41
After concealment using edge based interpolation	41.38	44.38	44.88

References

1. Y. Wang and Q.-F. Zhu, *Proc. of the IEEE* **86**, 974 (1998).
2. P. Yin, B. Liu, and H. H. Yu, *Proc. of. ICASSP* **3**, 1453 (2001).
3. S. Katzenbeisser and F. A. P. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House (2000).
4. D. L. Robie and R. M. Mersereau, *Proc. of ICIP*, 930 (2001).
5. W. Zeng, and B. Liu, *IEEE Transactions on Circuits and Systems for Video Technology* **9**, 648 (1999).

Combining MPEG-7 Based Visual Experts For Reaching Semantics

Medeni Soysal[1,2] and A. Aydin Alatan[1,2]

[1] Department of Electrical and Electronics Engineering, M.E.T.U.,
[2] TÜBİTAK BİLTEN,
Balgat, 06531, Ankara, Turkey
{medeni.soysal@bilten, alatan@eee}.metu.edu.tr

Abstract. Semantic classification of images using low-level features is a challenging problem. Combining experts with different classifier structures, trained by MPEG-7 low-level color and texture descriptors is examined as a solution alternative. For combining different classifiers and features, two advanced decision mechanisms are proposed, one of which enjoys a significant classification performance improvement. Simulations are conducted on 8 different visual semantic classes, resulting in accuracy improvements between 3.5-6.5%, when they are compared with the best performance of single classifier systems.

1 Introduction

Large collections of digital multimedia data are used in various areas today [1]. This is an inevitable result of the technological advances that make it easy to create, store and exchange digital multimedia content. Most of this content is indexed by manual annotation of data during the process of input. Although manual annotation is inevitable for some cases, replacing it with automatic annotation whenever possible, lifts a great burden.

MPEG-7 standard comes up with many features, supporting both manual and automatic annotation alternatives. In this standard, although many detailed media descriptions for manual annotation exist, automatic annotation is encouraged by many audio-visual low-level descriptors. These low-level descriptions (features) can be extracted automatically from the data by using many state-of-the-art algorithms. In this context, the most challenging problem is to find some relations between these low-level features and high-level semantic descriptions, desired by typical users. Focusing on visual descriptors, some typical examples of such high-level visual semantic descriptions can be indoor, sea, sky, crowd, etc.

Classification using low-level descriptions is widespread in many different areas as well as image classification. However, utilization of standardized features like those in MPEG-7 and combining them are relatively new ideas in the area of image classification. The work presented reaches one step beyond the previous approaches, and performs supervised classification of still images into semantic classes by utilizing multiple standard-based features and various classifier structures concurrently in two different settings.

These settings, namely advanced decision mechanisms that are proposed in this paper are compared against common single classifier-single descriptor setting and some other techniques in terms of classification performances. In this way, interesting and important relations are revealed.

The paper is organized as follows. In Section 2, low-level image features that are used in classification are introduced. Classifier types used and modifications on them are explained in Section 3. Various well-known basic methods to combine the experts, which use the features explained in Section 2 and have one of the classifier structures in Section 3, are discussed in Section 4. Two advanced decision mechanisms are developed in Section 5, as an alternative to the classical methods. These proposed decision mechanisms are compared with the best performance of single experts experimentally in Section 6. Section 7 summarizes the main results of the work and offers concluding remarks.

2 Low-Level Image Features

Successful image classification requires a good selection among low-level representations (i.e. features). In this research, color and texture descriptors of MPEG-7 [2] are utilized. A total of 4 descriptors are used, while two of them (color layout and color structure) are color-based, the other two (edge histogram and homogeneous texture) are texture descriptors.

MPEG-7 Color Layout descriptor is obtained by applying DCT transformation on the 2–D array of local representative colors in YCbCr space. Local representative colors are determined by dividing the image into 64 blocks and averaging 3 channels on these blocks. After DCT transformation, a nonlinear quantization is applied and first few coefficients are taken. In these experiments, only 6 coefficients for luminance and 3 coefficients for each chrominance are used, respectively [2].

MPEG-7 Color Structure descriptor specifies both color content (like color histogram) and the structure of this content by the help of a structure element [2]. This descriptor can distinguish between two images in which a given color is present in identical amounts, whereas the structure of the groups of pixels is different.

Spatial distribution of edges in an image is found out to be a useful texture feature for image classification [2]. The edge histogram descriptor in MPEG-7 represents local edge distribution in an image by dividing the image into 4x4 sub-images and generating a histogram from the edges present in each block. Edges in the image are categorized into five types, namely, vertical, horizontal, 45° diagonal, 135° diagonal and non-directional edges. In the end, a histogram with 16x5=80 bins is obtained, corresponding to a feature vector with 80 dimensions.

MPEG-7 Homogeneous Texture descriptor characterizes the region texture by mean energy and energy deviation from a set of frequency channels. The channels are modeled by Gabor functions and the 2-D frequency plane is portioned into 30 channels. In order to construct the descriptor, the mean and the standard deviation of the image in pixel domain is calculated and combined into a feature vector with the mean and energy deviation computed in each of the 30 frequency channels. As a result, a feature vector of 62 dimensions is extracted from each image [2].

3 Classifiers

In this research, 4 classifiers are utilized, which are Support Vector Machine [8], Nearest Mean, Bayesian Plug-In and K-nearest neighbors [4]. Binary classification is performed by experts obtained via training these classifiers with in-class and informative out-class samples. These classifiers are selected due to their distinct natures of modeling a distribution. For distance-based classifiers (i.e. Nearest Mean and K-Nearest Neighbor) special distance metrics compliant with the nature of the MPEG-7 descriptors are utilized. Since the outputs of the classifiers are to be used in combination, modifications are achieved on some of them to convert uncalibrated distance values to the calibrated probability values in the range [0,1]. All of these modifications are explained in detail along with the structure of the classifiers in the following subsections.

3.1. Support Vector Machine (SVM)

SVM performs classification between two classes by finding a decision surface via certain samples of the training set. SVM approach is different from most classifiers in a way that it handles the risk concept. Although other classical classifiers try to classify training set with minimal errors and therefore reduce the empirical risk, SVM can sacrifice from training set performance for being successful on yet-to-be-seen samples and therefore reduces structural risk [8]. Briefly, one can say that SVM constructs a decision surface between samples of two classes, maximizing the margin between them. In this case, a SVM with second-degree polynomial kernel is utilized. SVM classifies any test data by calculating the distance of samples from the decision surface with its sign signifying which side of the surface they reside.

On the other hand, in order to combine the classifier outputs, each classifier should produce calibrated posterior probability values. In order to obtain such an output, a simple logistic link function method, proposed by Wahba [5] is utilized as below.

$$P(\text{in-class} \mid x) = \frac{1}{1+e^{-f(x)}} \tag{1}$$

In this formula, $f(x)$ is the output of SVM, which is the distance of the input vector from the decision surface.

3.2. Nearest Mean Classifier

Nearest mean classifier calculates the centers of in-class and out-class training samples and then assigns the upcoming samples to the closest center. This classifier again, gives two distance values as output and should be modified to produce a posterior probability value. A common method used for K-NN classifiers is utilized in this case [6]. According to this method, distance values are mapped to posterior probabilities by the formula,

$$P(w_i \mid x) = \frac{1}{d_{m_i}} / \sum_{j=1}^{2} \frac{1}{d_{m_j}} \quad (2)$$

where d_{mi} and d_{mj} are distances from the i^{th} and j^{th} class means, respectively. In addition, a second measure recomputes the probability values below a given certainty threshold by using the formula [6]:

$$P(w_i \mid x) = \frac{N_i}{N} \quad (3)$$

where N_i is the number of in-class training samples whose distance to the mean is greater than x, and N is the total number of in-class samples. In this way, a more effective nearest mean classifier can be obtained.

3.3. Bayesian Gaussian Plug-In Classifier

This classifier fits multivariate normal densities to the distribution of the training data. Two class conditional densities representing in-class and out-class training data are obtained as a result of this process [4]. Bayesian decision rule is then utilized to find the probability of the input to be a member of the semantic class.

3.4. K-Nearest Neighbor Classifiers (K-NN)

K-NN classifiers are especially successful while capturing important boundary details that are too complex for all of the previously mentioned classifiers. Due to this property, they can model sparse and scattered distributions with a relatively high accuracy.

Generally, the output of these classifiers are converted to probability, except for K=1 case, with the following formula:

$$P(w_i \mid x) = K_i / K \quad (4)$$

where K_i shows the number of nearest neighbors from class-i and K is the total number of nearest neighbors, taken into consideration. This computation, although quite simple, underestimates an important point about the location of the test sample relative to in-class and out-class training samples. Therefore, instead of the above method, a more complex estimation is utilized in this research:

$$P(w_i \mid x) = \sum_{y_j} \frac{1}{d(x, y_j)} / \sum_{i=1}^{k} \frac{1}{d(x, y_i)} \quad (5)$$

where y_j shows in-class nearest neighbors of the input and y_i represent all k-nearest neighbors of the input.

Although, this estimation provides a more reliable probability output, it is observed that applying another measure to the test samples with probabilities obtained by (5) below a threshold also improves the result. This measure utilizes the relative positions of training data among each other [6]. This metric is the sum of the distances of each in-class training sample to its k in-class nearest neighbors:

$$g(x) = \sum_{i=1}^{k} d(x, y_i) \qquad y_i: i^{th} \text{ in-class nearest neighbor} \qquad (6)$$

After this value is computed for each training sample and input test sample, the final value is obtained by,

$$P(in-class \mid x) = 1 - (N_i / N) \qquad (7)$$

where N_i is the number of in-class training samples with g(x) value smaller than the input test sample and N is the number of all n-class training samples. In this way, a significant improvement is achieved in 3-NN, 5-NN, 7-NN and 9-NN classifier results.

For 1-NN case, since the conversion techniques explained here are not applicable, the probability estimation technique employed in the case of nearest mean classifier is applied.

4 Expert Combination Strategies

Combining *experts*, which are defined as the instances of classifiers with distinct natures working on distinct feature spaces, has been a popular research topic for years. Latest studies have provided mature and satisfying methods. In this research, six popular techniques, details of which are available in literature are adopted [3]. In all of these cases, a priori probabilities are assumed as 0.5 and the decision is made by the following formula:

$$P(in-class \mid X) = \frac{P_1}{P_1 + P_2} \qquad (8)$$

Here, P_1 is the combined output of experts about the likelihood of the sample X belonging to the semantic class while P_2 is the likelihood for X not belonging to the semantic class. Decision is made according to the Bayes' rule; if the likelihood is above 0.5, the sample is assigned as in-class, else out-class. P_1 and P_2 are obtained by using combination rules, namely, *product rule*, *sum rule*, *max rule*, *min rule*, *median rule* and *majority vote* [3]. In product rule, R experts are combined as follows,

$$P_1 = \prod_{i=1}^{R} P_i(in-class \mid X) \qquad P_2 = \prod_{i=1}^{R} P_i(out-class \mid X) \qquad (9)$$

Similarly, sum rule calculates the above probabilities as,

$$P_1 = \sum_{i=1}^{R} P_i(in-class \mid X) \qquad P_2 = \sum_{i=1}^{R} P_i(out-class \mid X) \qquad (10)$$

Others, which are derivations of these two rules, perform the same calculation as follows:

Max Rule $\qquad P_1 = \max_{i=1}^{R} P_i(in-class \mid X) \qquad P_2 = \max_{i=1}^{R} P_i(out-class \mid X) \qquad (11)$

Min Rule	$P_1 = \min_{i=1}^{R} P_i(in-class \mid X)$	$P_2 = \min_{i=1}^{R} P_i($
Median Rule	$P_1 = \underset{i=1}{\overset{R}{med}} P_i(in-class \mid X)$	$P_2 = \underset{i=1}{\overset{R}{med}} P_i($

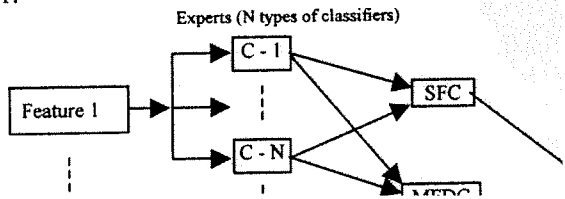
Lastly, majority vote (MV) counts the number of experts with higher than 0.5 and assigns to P_1. P_2 is the number of voting exp

Majority Vote	$P_1 = \frac{N_1}{N_1 + N_2}$	$P_2 = \frac{N_2}{N_1 + N_2}$	N_1 : # experts N_2 : # experts

5 Advanced Decision Mechanisms

In order to improve the classification performance, which is ach bination strategies, two different advanced mechanisms are mechanisms, namely Multiple Feature Direct Combination (M Feature Cascaded Combination (MFCC), use the output of sing ferent ways. They are applied only to semantic classes, for wh the low-level features are required. In these mechanisms, only are involved, leaving out 3-NN, 7-NN and 9-NN type experts, nance of K-NN. These experts are based on SVM, Nearest Mea Plug-in, 1-NN and 5-NN classifiers.

MFDC mechanism combines output of sole experts, which a level features, in a single step. For instance, 3x5=15 experts w class that is represented by three different low-level features. I Feature Combination (SFC) outputs are utilized. SFC combines same low-level feature and gives a single result. Next, MFCC u SFC to generate a resultant in-class probability. These two mecl in Figure 1.



6 Implementation Issues

A total of 1600 images, collected from various sources and having different resolutions, are used for training and test phases. A total of eight semantic classes are classified. For each class, 100 in-class and 100 out-class samples are used. Boosting [4] is used to prevent dependence of results on images. Five tests are performed by taking 20 distinct samples from each of in-class and out-class data sets, and training the experts by remaining classified data consisting of 80 in-class and 80 out-class training samples. Results are evaluated by considering the average of these five tests.

The semantic classes selected for these tests have the common property of being convenient to be inferred from low-level visual features extracted from the entire image. This means that, the characteristics of these classes are usually significant in the entire image and therefore the need for segmentation is mostly avoided.

Eight classes that are subjects of the tests are *football*, *indoor* (*outdoor*), *crowd*, *sunset-sunrise*, *sky*, *forest*, *sea* and *cityscape*. For each class, MPEG-7 color and texture descriptors that proved to capture the characteristics best in the pre-experiments, are utilized. The corresponding features to classifying these classes are tabulated in Table 1.

Table 1. Semantic classes and related features

Semantic Class	Low-Level Features
Football	Color Layout
Indoor	Edge Histogram
Crowd	Homogeneous Texture
Sunset-Sunrise	Color Layout, Color Structure, Edge Histogram
Sky	Color Layout, Color Structure, Homogeneous Texture
Forest	Color Structure, Edge Histogram, Homogeneous Texture
Sea	Color Layout, Homogeneous Texture
Cityscape	Color Structure, Edge Histogram, Homogeneous Texture

Table 2. Performances of SFC v.s. single experts

		Max Single	Single Feature Comb. (SFC)					
			Prd	Sum	Max	Min	Med	MV
Football	Accuracy	91.0	87.5	89.5	87.5	87.5	90.0	91.0
	Precision	91.6	98.8	98.8	97.3	97.3	98.8	92.7
	Recall	91.0	76.0	80.0	77.0	77.0	81.0	89.0
Indoor	Accuracy	83.0	84.0	83.0	83.5	83.5	81.0	84.0
	Precision	81.1	91.3	91.0	88.3	88.3	90.8	90.4
	Recall	87.0	75.0	73.0	77.0	77.0	69.0	76.0
Crowd	Accuracy	79.5	75.5	81.0	77.0	77.0	79.0	78.5
	Precision	83.5	72.5	81.8	73.3	73.3	81.6	79.6
	Recall	73.0	84.0	80.0	87.0	87.0	75.0	77.0

7 Experimental Results

Combination of experts has been tested on eight semantic classes. For the first three of these classes (*football*, *indoor* and *crowd*), only one representative low-level feature is used and therefore only Single Feature Combination (SFC) is available. Other five classes (*sunset-sunrise*, *sky*, *forest*, *sea* and *cityscape*) are represented by multiple features and therefore advanced decision mechanisms (MFDC and MFCC) are also applicable. Performances of the techniques on these two sets of classes are presented separately in different tables, Table 2 and Table 3, respectively. In order to provide a good basis of comparison, for each class, the result of an "optimal combination formula" which is obtained by combining experts with the best results, is also included. Obviously, such a case is not practical, since it should be determined case-by-case basis for each class.

In this section, although the accuracy results are used for comparison, precision and recall results are also included in the tables. This is because of the fact that they convey information about different properties of the techniques, which is hidden in accuracy.

Table 3. Performances of single experts, SFC, MFDC, and MFCC on different classes.

		Max Single	Max SFC	MFCC						MFDC						Optimal Comb. Formula	
				Prd	Sum	Max	Min	Med	MV	Prd	Sum	Max	Min	Med	MV		
Sunset Sunrise	Accuracy	92.5	92.0	92.5	90.0	92.0	92.0	90.0	90.0	92.5	91.0	84.5	91.0	91.0	91.0	93.5	Prd CSD-1NN CSD-NM
	Precision	90.9	88.8	93.5	91.2	92.6	92.6	91.2	91.2	93.5	93.1	82.2	91.4	90.6	90.6	92.5	
	Recall	95.0	97.0	92.0	89.0	92.0	92.0	89.0	89.0	92.0	89.0	91.0	91.0	92.0	92.0	95.0	
Sky	Accuracy	93.0	92.5	96.0	96.5	94.0	95.0	96.5	96.5	96.0	97.0	83.0	88.5	95.5	95.5	96.0	Sum CSD-SVM CSD-1NN HTD-5NN
	Precision	89.0	92.3	94.5	94.7	94.5	95.4	94.7	94.7	94.5	95.4	86.4	97.6	92.2	92.2	94.5	
	Recall	100.0	94.0	98.0	99.0	94.0	95.0	99.0	99.0	98.0	99.0	79.0	79.0	100.0	100.0	98.0	
Forest	Accuracy	79.0	82.0	86.5	86.0	85.0	85.0	85.5	85.5	86.5	84.5	78.0	83.5	83.0	83.0	85.0	Max CSD-5NN EHD-1NN HTD-SVM
	Precision	78.4	84.6	85.7	84.3	84.3	84.3	84.1	84.1	85.7	84.0	75.3	84.1	81.0	81.0	83.8	
	Recall	81.0	80.0	90.0	90.0	88.0	88.0	89.0	89.0	90.0	87.0	86.0	86.0	88.0	88.0	88.0	
Sea	Accuracy	80.5	83.0	86.0	86.0	86.0	86.0	86.0	60.0	86.0	84.5	74.0	79.0	82.5	81.0	85.5	Prd CLD-Med HTD-Max
	Precision	75.8	81.8	89.0	89.0	89.0	89.0	89.0	56.0	89.0	89.7	73.3	83.5	88.6	85.5	94.5	
	Recall	93.0	85.0	84.0	84.0	84.0	84.0	84.0	64.0	84.0	80.0	75.0	75.0	77.0	79.0	76.0	
Cityscape	Accuracy	82.0	81.5	85.0	86.5	82.0	82.0	87.0	87.0	85.0	83.5	71.0	77.0	81.0	81.5	85.5	Med CSD-SVM EHD-Bayes HTD-Bayes
	Precision	82.6	81.9	84.9	86.0	83.5	83.5	86.1	86.1	84.9	82.9	74.1	84.3	78.9	79.6	88.0	
	Recall	81.0	81.0	86.0	87.0	81.0	81.0	88.0	88.0	86.0	84.0	67.0	67.0	85.0	85.0	84.0	

For the classes in Table 2, it is seen that SFC leads with at least one rule except for the *football* case. However, improvements are not significant and also performance depends on the choice of the best combination for each of the above classes. For *football*, the majority vote rule gives the same result (% 91) with the best expert, which is a 1-NN. *Indoor* class is classified slightly better than the best expert (% 83) by product and majority vote results (% 84). In *crowd* classification, sum rule reached 81% and beat 9-NN classifier, whose performance was 79.5%.

Significant improvements are observed in the cases, where the proposed advanced decision mechanisms are applicable. MFDC and MFCC outperform the best single expert and best SFC for nearly all classes. The only case in which advanced decision mechanisms do not yield better results than the best single expert is *sunset (sunrise)* classification.

MFDC though being successful against single experts, could not beat the "optimal combination formula" in most of the cases. However, the "optimal combination formula" gives inferior results against MFCC for the most cases. For instance, MFCC improves the performance of classifications, especially when its second stage combination rule is fixed to median, while SFCs in the previous stage are obtained by the product rule. This should be due to the fact that these two rules have properties, which compensate the weak representations of each other. Product rule, although known to have many favorable properties, is a "severe" rule, since a single expert can inhibit the positive decision of all the others by outputting a close to zero probability [3]. Median rule, however, can be viewed as a robust average of all experts and is therefore more resilient to this weakness belonging to the product rule. This leads us to the observation that combining the product rule and the median rule is an effective method of increasing the modeling performance. This observation on MFCC is also supported by a performance improvement of 3.5% for *sky*, 6.5% for *forest*, 5.5% for *sea* and 5% for *cityscape* classification, when it is compared against the best single classifier. MFCC also achieves a performance improvement of at least 1-2% over even the manually selected "optimal combination formula".

Another important fact about the performances achieved in classification of these classes using advanced decision mechanisms is the increase in precision values they provide. In the application of classification of these methods to large databases with higher variation compared with data sets used in experiments, usually recall values are sustained, however precision values drop severely. The methods proposed in this text, therefore have also an effect of increasing robustness of classification.

In addition, although the averages of the test sets are displayed for each class, when the separate test set performances are analyzed, MFCC shows quite stable characteristics. The variance of its performance from one test set to another is less than all others. Typical classification results can also be observed at our ongoing MPEG-7 compliant multimedia management system site, `BilVMS` (http://vms.bilten.metu.edu.tr/).

8 Conclusion

Reaching semantic information from low-level features is a challenging problem. Most of the time, it is not enough to train a single type of classifier with a single low-level feature to define a semantic class. Either it is required to use multiple features to represent the class, or it is needed to combine different classifiers to fit a distribution to the members of the class in the selected feature space.

Advanced decision mechanisms are proposed in this paper, and among the two methods, especially, Multiple Feature Cascaded Combination (MFCC) achieves significant improvements, even in the cases where single experts have already had very high accuracies. The main reason for this improvement is the reliability and stability the combination gains, since experts that are good at modeling different parts of the class distribution are combined to complement each other. For MFCC, it is observed that classification performance significantly improves, when correct combination rules are selected at each stage. For instance, combining the product rule results of the first stage by using median rule is found out to be quite successful in all cases. This observation can be explained by the complementary nature of the rules.

References

1. Forsyth, D.A.: Benchmarks for Storage and Retrieval in Multimedia Databases. Proc. Of SPIE, Vol. 4676 SPIE Press, San Jose, California (2002) 240-247
2. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7. John Wiley&Sons Ltd. England (2002)
3. Kittler, J., Hataf, M., Duin, R.P.W., Matas, J.: On Combining Classifiers. IEEE Trans. PAMI Vol. 20. No. 3. Mar. 1998.(1998) 226-239
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley&Sons Ltd. Canada (2001)
5. Platt, J.C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers. MIT Press. Cambridge. MA (1999)
6. Arlandis, J., Perez-Cortes, J.C., Cano, J.: Rejection Strategies and Confidence Measures for a k-NN Classifier in an OCR Task. IEEE (2002)
7. Tong, S., Chang, E.: Support Vector Machine Active Learning for Image Retrieval. Proc. ACM. Int. Conf. on Multimedia. New York (2001) 107-118
8. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag. New York (1995)

Summarizing Video: Content, Features & Hmm Topologies

Yağız Yaşaroğlu[1,2] and A. Aydın Alatan[1,2]

[1] Department of Electrical and Electronics Engineering, M.E.T.U.,
[2] TÜBİTAK BİLTEN,

Balgat, 06531, Ankara, TURKEY
{yagiz.yasaroglu@bilten, alatan@eee}.metu.edu.tr

Abstract. An algorithm is proposed for automatic summarization of multimedia content by segmenting digital video into semantic scenes using HMMs. Various multi-modal low-level features are extracted to determine state transitions in HMMs for summarization. Advantage of using different model topologies and observation sets in order to segment different content types is emphasized and verified by simulations. Performance of the proposed algorithm is also compared with a deterministic scene segmentation method. A better performance is observed due to the flexibility of HMMs in modeling different content types.

1 Introduction

The most critical issue in multimedia management is automation. Rapid growth of multimedia content requires sophisticated analysis algorithms running with minimum user intervention. However, automatically extractable clues are usually low-level descriptions of the multimedia content (e.g. color, shape, pitch), and relating them to high-level semantic descriptions (e.g. dialogue, car, Beethoven) is a difficult problem to solve. An example is automatic analysis and extraction of the scene structure in digital video. Such an analysis is valuable, since it provides better indexing and more concise summaries of videos, compared to shot-based summarization techniques.

Apart from the sheer size of data that needs to be analyzed, another problem is the diversity of content types available. It is obvious that a soccer video does not have much in common with a documentary video from the summarization point of view. Their production styles, production purposes and properties of the end result are different. Moreover, sometimes examples within the same genre are not similar enough to be analyzed successfully using the same method (e.g. movies might be produced by directors with different styles). Thus, it is virtually impossible to build an automatic algorithm that successfully analyzes all different content types. Since different content types are generated using different processes, different models are needed to analyze them. Likewise, characteristic properties of particular content types are different, which requires the use of different low-level descriptions of the content.

In this paper, different video content types are analyzed by automatically extracting various low-level properties and summarized using a Hidden Markov Model (HMM), which takes the low-level properties as observation inputs. In the following sections, two different content types are analyzed using different combinations of low-level observations and HMM topologies. Finally, the HMM-based system is also compared with a deterministic approach explained in [1].

Throughout this paper, a 'shot' is used to mean a continuous recording of a camera, whereas a 'scene' means a temporally adjacent and semantically meaningful collection of shots.

2 Video Summarization

Current trend in video summarization can be broadly classified into two major classes, as model-based or similarity-based (clustering) approaches. Model-based approaches, as their name implies, either try to match a model from a library to the observed data for classification [2-4], or utilize a model to segment the given data [5,6]. In these approaches, the models can be either finite-state machines [2], or HMMs [3-6]. On the other hand, clustering-based approaches [1,7] do not rely on any model, but use a similarity measure between visual clues.

A popular approach in video summarization is classifying particular scene types in a video stream. For example, a finite state machine that represents the structure of dialogue and action scenes involving two parties can be used to summarize story-based video sequences [2]. In [3], authors develop HMMs to define play and break scenes in a soccer video and use dynamic programming for parsing. A similar study, detects highlights in baseball videos by developing Hidden Markov Models that represent different scene types in a baseball sequence [4].

On the other hand, some methods take into account the global structure of the video and segment the data based on this structure. In [5], different HMMs that model documentary structures are presented, whereas in [6], dialogue scenes in story-based videos are segmented using HMMs that imitate the inherent grammar of videos.

As a different approach, a clustering method [1] uses histogram-based visual similarity and activity similarity measures to cluster shots in a video into semantic classes. The temporal relationships of shots are also taken into account by merging a shot into a scene based on distance between shots in time. An intelligent rule-based post-processing method further refines the clustering method, effectively merging temporally interleaved scenes [1]. A similar method [7] employs color, edge, shape, audio and close caption features, while using Bayesian Belief Networks to extract topics of program portions in a video.

3 Hidden Markov Models

Hidden Markov Models (HMM) are powerful statistical tools that have been successfully utilized in speech recognition and speaker identification fields [8]. They have

also found applications in content-based video indexing area for solving video scene segmentation [4,5,6,9]. Recently, other researchers approached modified HMM structures such as Coupled HMMs [10] or used HMMs in conjunction with other probabilistic methods, such as dynamic programming [3], in alternative video segmentation schemes. The most critical design issues for HMM-based modeling are defining the hidden states, finding state topology and deciding the observable symbols at each state. After these initial design steps, determining the statistical parameters (Baum-Welch algorithm [8]) and finding a state-sequence for an input (Viterbi algorithm [8]) have well-known solutions.

For video scene modeling, assigning scenes of the content to states of the HMM is the most straightforward approach. According to the scene classifications of content producers, such as *establishing scene*, *dialogue scene*, etc., the states of the model can be determined. HMM states can be connected to each other using different HMM topologies. One possibility is a left-to-right HMM state topology, but this option is not found feasible, since the number of scenes is not known beforehand [6]. In this paper, circular HMM topologies with different number of states are examined (Fig. 1). Section 4 explains the final critical design issue, the choice of observable output symbols of HMM at each hidden state.

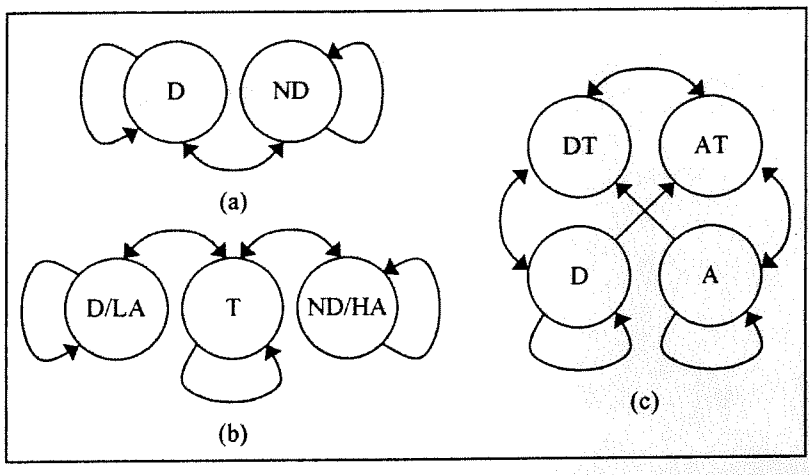


Fig. 1. Different HMM topologies. (D: Dialogue, ND: Non-dialogue, LA: Low-action, HA: High-action, T: Transition, A: Action, DT: Dialogue transition, AT: Action Transition)

4 Content Types for Summarization

Two different content types with different characteristic properties are defined.

4.1 Type I: Dialogue-Driven Content

Story-based dialogue-driven video content is classified as Content Type I. Videos of this type are made up mainly of dialogue scenes following each other to build a story. Situation comedies, dramas, and some TV series fall into this category. Motion activity feature is not expected to be of much use in segmentation of videos belonging to this content type, since dialogues typically have consistent, low motion activity values. On the other hand, presence of speech and face gives important information about the scene structure, since they are suitable for dialogue segmentation, as demonstrated in [6].

4.2 Type II: Action-Driven Content

Similar to Type I, video belonging to Content Type II is also story-based, but they are action-driven. Action scenes are at least as important as dialogue scenes, and story is presented through a sequence of dialogue and action scenes. All kinds of action movies (thrillers, sci-fi, detective, etc.) fall into this category. The motion activity feature is expected to perform well with this content type, since high-action scenes and low-action scenes exhibit different motion activity properties.

5 Automatic Summarization System

For all different content types, features and topologies, HMM-based video summarization system should consist of three stages: Pre-processing, feature extraction and decision-making (Fig.2).

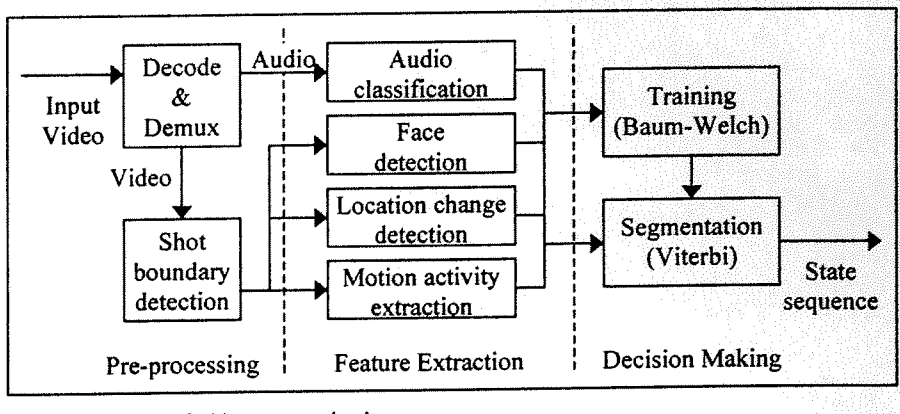


Fig. 2. HMM-based video summarization

In the preprocessing stage, compressed video stream should be decoded and demultiplexed for multi-modal analysis, and it should be further parsed into its shots. In

the feature extraction stage, audio stream might be analyzed to identify the shots having silence, speech or music content; and video stream can be analyzed to detect shots containing faces, location changes and to find motion activities for each shot. It is important to note that, depending on the type of the video, not all of the features might be needed in segmentation, and some of the features may not be suitable for segmenting certain content types.

The HMM comprising the decision-making stage depends on type of the video being segmented. By using an HMM, the system is trying to model the process by which observed features are generated. Therefore models for different types of videos should also be different.

Output of the decision making stage is a sequence of states in which each state corresponds to a shot in the input video and attaches a type label to it. Following the decision making stage, consecutive shots of the same type are merged together into scenes. In this merging process, a scene is assumed to be composed of one or more shots of type ND, T, DT or AT, followed by one or more shots of type D, A, HA or LA. This approach enables videos to be segmented seamlessly into scenes. It is trivial to compose a visual summary by selecting key frames from each scene after this point.

5.1 HMM Training

HMM training has well-defined solutions [8], however it requires special attention from video segmentation point of view. Although the video content usually shows similar properties within genres (e.g. dialogues consist of alternating shots of speaking people or a goal is immediately followed by slowing down of action in soccer), some higher-level, subtler properties can be different. These high-level features are usually based on director's choices or the script. For example, even if one compares two movies of the same genre, assuming their directors and scripts are different, they may exhibit different frequency of action scenes, different dialogue lengths, different usage of camera, etc. Fortunately, a trained HMM is able to capture all these properties. Hence, self-training the model with the input video allows better modeling of the underlying process by which the video is generated, and thus provides a better performance. In fact, this process itself is not "training", but simply a model parameter extraction.

In the proposed system, a fixed initial model is used for videos of a content type. After feature extraction, the model is trained with extracted features, and at last, the movie is parsed using self-trained HMM. The initial model parameters for HMMs are obtained according to experience, and shown to be stable initial models. During simulations, up to 20% variations on the initial observation probabilities converge to the same trained models, as long as transition probabilities are kept constant.

5.2 Extracted Features

The fundamental low-level features are standardized in MPEG-7. The utilized descriptors from this standard are face, color histogram, motion activity and audio parameters. Each shot in the video sequence is labeled by these descriptors.

Face Detection: Presence of faces is a clue for dialogues and so ability to detect faces is valuable for video scene segmentation. Face detection is a quite mature topic with diverse solutions [11]. Most of these approaches are based on the simple fact that human skin color occupies a very narrow region in any 3-D color space. Hence, the segmentation of image points belonging to this region in the color space gives a good initial estimate of the skin-colored regions. YUV color space has been used in the system with simple heuristics [6]. Sample frames are taken from each shot periodically, are analyzed for existence of faces, and results are voted within each shot.

Audio Analysis: Humans can understand the scene structure of a video by only listening to its audio content. Even the semantically lower level properties of the audio track (e.g. the presence of speech or music) are valuable for segmentation. Audio track is segmented into three classes as silence, speech and music [6]. The segmentation process begins with calculating the energy of audio segments. Low energy segments are labeled as silence. High-energy segments are checked for periodicity using an autocorrelation function. Since both voiced sounds and music may have significant peaks in their autocorrelation function, Zero Crossing Rate (ZCR) [12] of these signals is also measured. ZCR detects abrupt changes that should occur in speech signals due to existence of both voiced (low ZCR) and unvoiced (high ZCR) sounds. Music signals are detected by a significant periodicity with small changes in ZCR. Shots are labeled according to their audio content type that has the longest duration.

Location Change Analysis: Location changes in a story-based video usually carry important information concerning scene boundaries. Especially on dialogue-based videos and on dialogue scenes in mixed-type videos scene boundaries tend to coincide with location changes. Mostly, a scene in a particular location consists of alternating shots of people or objects involved in the scene, which may be preceded and followed by wide shots of the location.

The problem of detecting location changes is approached by a windowed histogram comparison method. A fixed number of histograms are sampled from each shot within a temporal window, and mean and deviation histograms are calculated using these samples. As the window moves in time one sample at a time, similarity between the mean histogram and histogram of the sample at the front of the window is calculated. This similarity is compared with a deviation-dependent threshold to determine if there is a location change on that sample. If the number of location change samples exceeds that of other samples within a shot, the shot is labeled as a location change.

Motion Activity Analysis: Motion activity can be used in segmentation of Type II videos, since the low-activity scenes (e.g. dialogue scenes) and high-activity scenes exhibit contrasting activity behavior. Simulations on sample videos showed that low-activity scenes consistently tend to have low object motion values, whereas in high-activity scenes motion activity has variation, spanning all possible values. Motion activity information is extracted using the frame motion vectors [13]. The variance of magnitudes of these vectors is calculated for each frame, and variances are averaged for each shot. The results are quantized to 5 levels.

6 Simulations

Simulations are conducted on two phases. In the first phase, samples of both content types are segmented using different HMM topologies and observation sets. In the second phase, performance of the HMM based method is compared to that of a deterministic method [1].

Throughout the simulations, 4 videos recorded from a TV station are used. Two of the videos belong to Content Type I (a sitcom and a TV series) whereas the others belong to Content Type II (two Hollywood family movies). Non-story portions of the videos (commercial brakes, credits, summaries, etc.) are edited out before analysis and the video's ground truths are obtained for performance evaluation.

6.1 Content Type Simulations

First of all, audio and face features are used to segment the video into dialogue and non-dialogue shots. More than 2-state topologies are not used, since more state (scene) types do not have any semantic meaning. The results (Table 1) indicate that audio and face features alone are not successful in segmentation of the sample set. Closer examination of the results reveals that the videos are highly under-segmented.

Table 1. Recall / precision values for different HMM topologies using audio and face features

	Audio, face
Recall / precision	2-state
Type I	0.139 / 1.000
Type II	0.450 / 1.000

The next experiment adds the location change feature to the system, and this time 3- and 4-state topologies are used as well, since location changes imply transitions. The results in Table 2 show that dialogue-driven (Type I) content is segmented quite well with this set of features, using the 2-state topology. Generally Content Type I is better segmented than Content Type II with this feature set.

Table 2. Recall / precision values for different HMM topologies using audio, face and location change features

Audio, face, location change			
Recall / precision	2-state	3-state	4-state
Type I	1.000 / 0.742	0.778 / 0.489	0.584 / 0.548
Type II	0.741 / 0.784	0.584 / 0.438	0.800 / 0.648

After adding the motion activity feature to the observation set, the performance decreases (Table 3). Scenes are observed to be over-segmented, although Content Type I still has better results.

Table 3. Recall / precision values for different HMM topologies using audio, face, location change and motion activity features

Audio, face, location change, motion activity		
Recall / precision	3-state	4-state
Type I	0.750 / 0.430	0.611 / 0.389
Type II	0.416 / 0.122	0.458 / 0.198

The final experiment in this phase involves motion activity and location change features, and topologies with more than two states. As observed in Table 4, this time content of Type II is segmented with 95% recall and 80% precision using a 3-state topology.

Table 4. Recall / precision values for different HMM topologies using location change and motion activity features

Location change, motion activity		
Recall / precision	3-state	4-state
Type I	0.694 / 0.363	0.806 / 0.568
Type II	0.950 / 0.800	0.742 / 0.800

These results indicate important conclusions. The second observation set (audio, face and location change) is suitable for segmentation of Type I videos. This is expected, since Type I video are made up mainly of dialogue scenes (face, speech and no-change), and non-dialogue shots (not face and speech), or location change shots acting as scene boundaries. On the other hand, the fourth observation set (location change and motion activity) is observed to be suitable for segmentation of Type II video, since they are comprised of scenes having different motion activity content. For this case, location change shots act as transition scenes and rest of the video is segmented into high-action and low-action scenes. The final point to emphasize is utilization of all features for segmentation degrades performance.

Figure 3 shows two sample results from both content types. Video sequences are segmented into their scenes. The graphs show consecutive scenes in different colors. Scene boundaries (i.e. points at which colors change) are relevant, colors of individual scenes are not.

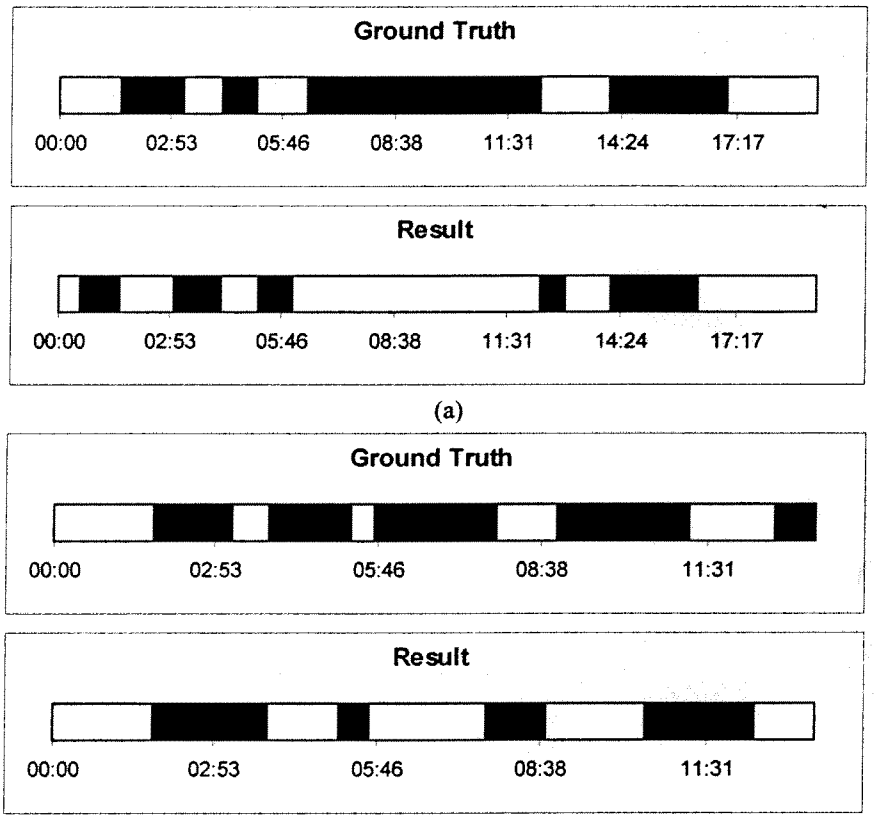


Fig. 3. Sample simulation results a) Dialog-driven (Type I), b) Action-driven (Type II). Consecutive scenes are colored differently; the actual colors (black or white) are irrelevant.

6.2 Comparison with Deterministic Approaches

The performance of the popular "semantic-level table of content construction technique" [1] is shown in Table 5. For the HMM-based method, the best performances for different content types are tabulated. For Type I content, audio, face and location change features are used with 2-state HMM topology, whereas for Type II content, location change and motion activity features on 3-state topology are utilized.

Table 5. Rule-Based vs. best performance HMM results

Recall / precision	Deterministic [1]	HMM for Type I	HMM for Type II
Type I	0.806 / 0.410	1.000 / 0.742	0.694 / 0.363
Type II	0.734 / 0.764	0.741 / 0.784	0.950 / 0.800

Results show that HMM-based method outperforms deterministic method with both types. Using different models for different content types is due to the flexibility of HMM-based strategy, which is not possible for deterministic approaches.

7 Conclusions and Future Work

An automatic HMM-based video segmentation scheme is presented. The method extracts four low-level features and generates scene structure information from these features. The necessity of different model topologies and observation sets for segmenting different content types is emphasized and verified through simulations conducted on two sample content types. Proposed method is also compared with an existing deterministic approach [1] and enjoyed a higher performance.

More models and features will be added to the system, and the possibility of automatically detecting a video's content type will be investigated. Modeling segments of videos instead of modeling entire videos will be evaluated as a new path to follow.

References

1. Rui, Y., Huang, T.S., Mehrotra, S: Constructing Table-of-Content for Videos. Multimedia Systems, Special section on Video Libraries 7 (1999) 359-368
2. Chen, L., T. Özsu: Rule-Based Scene Extraction from Video. Proc. of ICIP'02, Volv. 2 (2002) 737-740
3. Xie, L., Chang, S.-F., Divakaran A., Sun, H.: Structure Analysis of Soccer Video With Hidden Markov Models. Proceedings of ICASSP'02, Vol. 4 (2002) 1096-1099
4. Chang, P., Han, M., Gong, Y.: Extract Highlights from Baseball Game Video With Hidden Markov Models. Proc. of ICIP'02, Vol. 1 (2002) 609-612
5. Liu, T., Kender, J.R.: A HMM Approach to the Structure of Documentaries. CBAIVL'00 (2000) 111-115
6. Alatan, A.A., Akansu, A.N., Wolf, W.: Multi-Modal Dialogue Scene Detection using Hidden Markov Models for Content-based Multimedia Indexing. Int. Journal on Multimedia Tools and Applications, Kluwer Ac. (2001)
7. Jasinschi, R.S., et.al.: Video Scouting: An Architecture and System for the Integration of Multimedia Information. Proc. of ICASSP'01, Vol. 3 (2001) 1405-1408
8. Rabiner, L.R., Juang, B.-H.: Fundementals of Speech Recognition. Prentice Hall, Englewood, NJ, USA (1993)
9. Wolf, W.: Hidden Markov Model Parsing of Video Programs. Proc. of ICASSP'97 (1997), 2609-2611
10. Chu, S.M., Huang, T.S.: Audio-Visual Speech Modeling Using Coupled Hidden Markov Models. Proceedings of ICIP'02, Vol. 2, (2002) 2009-2012
11. Yang, M.-H., Kreigman, D.J., Ahuja, N.: Detecting Faces in Images. IEEE Trans. on PAMI, Vol. 24, (2002) 34-58
12. Saraceno C., Leonardi, R.: Identification of Story Units in Audio-Visual Sequences by Joint Audio and Video Processing,. Proc. of ICIP'98 (1998) 363-367
13. Peker, K.A., Divakaran, A., Papathomas, T.V.: Automatic Measurement of Intensity of Motion Activity of Video Segments. SPIE Conference on Storage and Retrieval for Media Databases, Vol. 4315 (2001) 341-351

UTILIZATION OF TEXTURE, CONTRAST AND COLOR HOMOGENEITY FOR DETECTING AND RECOGNIZING TEXT FROM VIDEO FRAMES

Serhat Tekinalp and A. Aydın Alatan
Department of Electrical and Electronics Engineering
M.E.T.U., Balgat, 06533, Ankara, TURKEY
e-mail: alatan@eee.metu.edu.tr[1]

ABSTRACT

It is possible to index and manage large video archives in a more efficient manner by detecting and recognizing text within video frames. There are some inherent properties of videotext, such as distinguishing texture, higher contrast against background, and uniform color, making it detectable. By employing these properties, it is possible to detect text regions and binarize the image for character recognition. In this paper, a complete framework for detection and recognition of videotext is presented. The results from Gabor-based texture analysis, contrast-based segmentation and color homogeneity are merged to obtain minimum number of candidate regions before binarization. The performance of the system is tested for its recognition rate for various combinations and it is observed that the results give recognition rates, reasonable for most practical purposes.

1. INTRODUCTION

Content-based information retrieval from digital video databases and media archives is a challenging problem and is rapidly gaining widespread research and commercial interest. In order to be able to index video sequences for retrieval, it is necessary to produce the natural language representation of the sequence automatically. This representation can most directly be extracted from information carriers, such as voice and text, but annotational information often appears only in image text. Such annotations are usually of readable quality, and may contain keywords, facilitate indexing.

The use of textual information as a key for an indexing system requires conversion of videotext present in image to ASCII form, which is commonly known as Optical Character Recognition (OCR) process. Unfortunately, commercially available OCR systems are designed for recognition of characters in printed documents [1]. The characters in these documents appear in uniform color on a clean background and the document is typically scanned in high resolution. Therefore, to be able to employ OCR for videotext requires segmentation of characters from background, which is usually arbitrarily complex.

For OCR systems to recognize characters efficiently, a video frame should be converted into a black and white image, where pixels on the characters of the videotext are black and the remaining pixels are white. This operation is not straightforward, since character pixels are often composed of a range of intensities and the boundary between the character region and the background is not perfectly defined. Hence, the binarization process should utilize *all* the distinctive properties of videotext, such as texture, high contrast, uniform color and constrained size. In this manuscript, a novel algorithm is proposed by utilizing all these properties and is fully tested including final character recognition stage.

2. RELATED WORK

The current research can be mainly divided into two major classes: connected-component-based approaches [5,6,7], and texture-based methods [3,4,8,9]. While the former tries to find characters as closed regions containing uniform color, the latter considers the text as a special class of texture. The boundary between these two methods is not very clear and there are also some hybrid approaches [2] that utilize both of these approaches.

Connected component-based approaches require high contrast and resolution in order to give an acceptable performance. Neighboring regions with similar color values, i.e. low contrast, may usually yield erroneous segmentation of the text [5,6,10]. In some methods, the assumption of having monochrome text or, a monochrome background surrounding the text is another limiting factor for the general performance [7]. Moreover, finding characters as a single closed region is only possible by utilizing high-resolution images [2]. However, for video indexing, such constraints are far not practical.

Among various texture-based text detection algorithms, the main difference is due to their representation of the texture. The main idea behind texture analysis is

This research is supported by TÜBİTAK BİLTEN

quantifying the distribution of different edge orientations. Among different models, Gabor filters with various orientations [8] or simple edge detectors for finding a specific direction [4] or multiresolution wavelet filtering [3,9] can be used for analyzing texture.

Existing work on videotext detection and recognition generally suffers from lack of robustness. Connected component-based techniques are more general in the sense that texture-based methods are usually based on assumptions about videotext character size. On the other hand, texture-based methods offer more reliability. Finally, the performance of these methods in practical applications is still not clear, since most of the proposed algorithms are not tested using quantitative character recognition results after the OCR stage.

3. VIDEOTEXT DETECTION

The binarization process for videotext extraction should ideally remove all image components other than videotext characters. Since the problem cannot be solved at global or pixel level, it is necessary to employ regional and object level image analysis [11].

Candidate regions are the areas of interest that have the possibility of containing text. The essence of the problem is finding the candidate regions efficiently by the help of all available information. In order to produce the binary image, which will be fed to the OCR for recognition, all candidate regions should be binarized locally with an appropriate threshold value specific to that region. The whole process can be summarized as in Fig. 1.

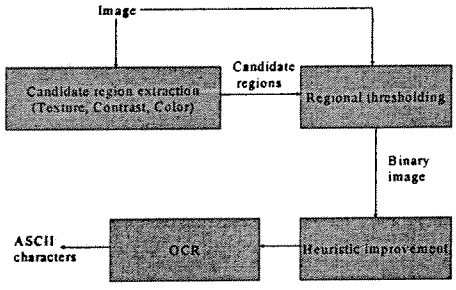


Figure 1: Block diagram overview of videotext detection and recognition process.

In order to find the candidate regions before binarization, texture, contrast and color-based segmentation results are obtained and they are appropriately merged to determine final set of candidate regions.

3.1 Texture Analysis for Candidate Region Extraction

Gabor filters have been one of the major tools for texture analysis [12,13] and also been utilized in videotext detection [8]. This technique has the advantage of analyzing texture in an unlimited number of directions and scales. This flexibility is very useful for videotext detection due to appearance of the videotext character edges in a diverse range of directions [8].

In our proposed method, a set of Gabor filters with a single scale and 8 different orientations are used to produce an eight dimensional feature vector for each 16x16 block of the image. A pre-trained feedforward neural network classifies this vector as text or non-text. In Fig. 2, a typical example video frame with its neural network output is shown.

Figure 2: A typical Gabor-based texture analysis result.

3.2 Contrast-based Candidate Region Segmentation

Another important distinctive property of videotext regions is their high contrast against background. For contrast analysis, a simple contrast measure proposed by Lienhart et. al. [6] is used. In this method, a binary contrast image is derived for each video frame. The absolute local contrast at position $I(x,y)$ is measured by

$$C(x,y) = \sum_{k=-r}^{r}\sum_{l=-r}^{r} G_{k,l}\left|I(x,y) - I(x-k,y-l)\right|$$

where $G_{k,l}$ denotes a 2D Gaussian smoothing filter, and r denotes the size of the local neighborhood. The value obtained by this measure is thresholded to get rid of the low contrast regions. Using the contrast image, the pixels, which show sufficiently high absolute local contrast, are marked with 1 whereas the rest is marked as 0. Fig.3 shows a contrast image and its thresholded version.

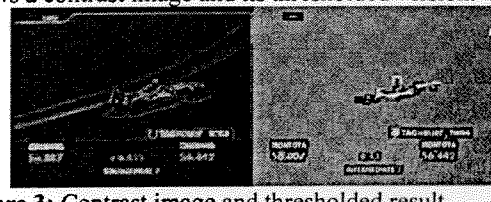


Figure 3: Contrast image and thresholded result.

3.3 Homogenous Color Region Segmentation

Segmenting the image into homogenous color regions is necessary in order to eliminate a set of regions from being a candidate to videotext areas. Hence, the image is decomposed into non-overlapping regions with color homogeneity. Although, many different methods can be utilized, a simple criterion is used to group pixels whose absolute gray level differences do not exceed δ within a region (after simulations, δ is selected as 20). After this recursive segmentation, regions, which cannot represent a character based on their areas, are marked as background and removed. In this approach, a region is identified as non-videotext and removed, if its height is greater than

1/8 of image height or its width is greater than 1/8 of image width. In Fig. 4, a typical video frame and the extracted candidate regions using region analysis are presented.

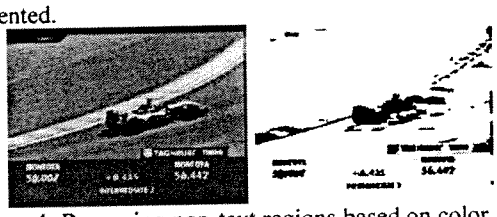


Figure 4: Removing non-text regions based on color

3.4. Regional Thresholding

The idea of binarization with regional thresholding is originally proposed by Dorai et.al. [10]. In their work, the candidate regions are first extracted by simple region analysis, but this approach suffers from the high amount of non-character regions, which cannot be eliminated by any heuristics or shape analysis. By combining texture and contrast-based videotext detection, which narrow regions of interest to text box level to support region analysis, limitations of region analysis may be eliminated. In our approach, iterative thresholding and boundary enhancement [10] methods are applied consecutively to each candidate region. While iterative thresholding calculates the threshold between two regions by averaging the mean intensities using $T=(\mu_1+\mu_2)/2$, the boundary enhancement method uses an average of intensities in the circumscribing boundary and inside the region.

$$T = \left(\sum I_{cb} + \sum I_i \right) \Big/ (N_{cb} + N_i)$$

After simulations, the results favor the consecutive utilization of these two methods.

4. EXPERIMENTAL RESULTS

Since the optimal combination between different candidate region detection strategies is not known, the simulations determine the best way to merge texture, contrast and color-based region segmentation results. The whole process is illustrated in Fig. 5. The input image, Fig.5.(a), is analyzed for its texture and color homogeneity in Fig.5.(b) and Fig.5.(c), respectively, in parallel (this approach is just for the particular example; contrast analysis may also be added). The intersection of the outputs of these parallel threads forms the input for the thresholding step, Fig.5.(d). After thresholding, a mask image is obtained in Fig.5.(e) where the candidate regions are formed by grouping connected pixels above a threshold. After thresholding, simple heuristics, such as vertical character alignment, is applied to image and binary image in Fig.5.(f) is formed before the OCR step.

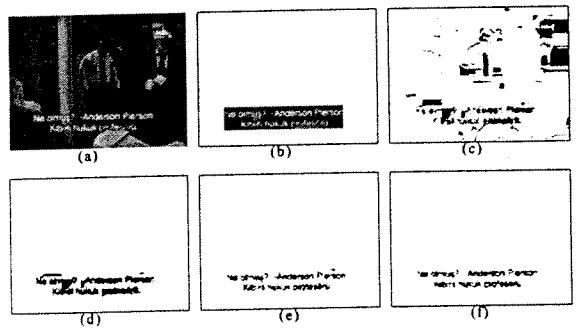


Figure 5: Pre-OCR steps : (a) original image (b) texture analysis output (c) region analysis output (d) Pre-thresholding output (e) Thresholding output (f) Heuristic improvement output

The experimental set is composed of 30 typical video frames of size 352x288, containing 964 characters. During experiments, evaluation of videotext binarization alternatives is based on both subjective and objective measures. Following the approach in [10], a subjective measure *binarization rate* is obtained by simply counting the number of "readable" binary characters inside the final mask. The OCR-based character *recognition rate* is also determined by using an commercial OCR (*ABBYY FineReader 5.0*) to obtain the most objective measure for the performance. None of the previous videotext detection methods has published their performance with an OCR. Summary of the results for videotext recognition experiments are given Table 1.

Candidate region detection	*Binarization rate*	*Recognition rate*
-	-	%25
R.A. + C.A.	%84	%57
R.A. + T.A.	%91	%59
R.A. + C.A + T.A.	%81	%56
R.A. + P.T.B.D.	%97	%66

Table 1: Summary of experimental results: R.A.:Region Analysis, C.A.:Contrast Analysis, T.A.:Texture Analysis, P.T.B.D.: Perfect (manual) Text Box Detection.

As shown in Table 1, different combinations of algorithms result in various recognition rates. The best performance is achieved when region analysis and texture analysis are used in parallel before binarization. A similar result, in fact the second maximum recognition rate, is achieved when region analysis is used in parallel with contrast analysis. In the last experiment, the text boxes are placed manually instead of automatic analysis and an upper bound for the performance is obtained. The

performance of the resulting system can be examined at *BilVMS* video management system [14].

5. DISCUSSION AND CONCLUSIONS

Throughout the review of the related work, it is observed that none of the proposed methods are mature enough to be accepted as a standard framework. However, as the literature suggests, the use of distinctive properties of videotext leads to a relatively high accuracy in detection of videotext character regions. This work is also concentrated on the use of these distinctive properties.

By employing contrast analysis and/or texture analysis, one can narrow the regions of interest, i.e. candidate character regions to textbox level. Additional analysis of homogeneous regions, further resolves the regions even to character level, since large homogeneous regions outside the textbox are usually connected to the regions between characters of videotext. The use of region analysis alone, however, does not give satisfactory results, since there usually exists a lot of non-character homogeneous regions, most of which cannot be eliminated by shape analysis or any heuristics. Therefore, region analysis should be supported with texture and/or contrast analysis in order to reduce the candidate regions to characters of videotext.

As can be seen from the experimental results, the use of contrast analysis with texture analysis together, decreases the recognition rate with respect to texture or contrast cases alone. This is due to the fact that both methods carry similar information and have drawbacks that limit the success of recognition. It is possible to conclude that contrast analysis and texture analysis techniques should be considered as alternatives to each other. The drawbacks of texture analysis are the limited block size and need for supervised training, whereas the main drawback of contrast analysis is the "over the average contrast videotext region" assumption. On the other hand, texture analysis is much more accurate than contrast analysis and contrast analysis is computationally much simpler than texture analysis.

The proposed system is quite successful for the images where a uniform colored frame surrounds the videotext and/or the characters of videotext are large and have high contrast with the surrounding background. As the character size reduces and the background complexity increases the recognition rate decreases. The overall performance of the system can be highly dependent on the low resolution of images, since the OCR utilized during simulations (as well as many other OCRs), requires high resolution input for accurate detection.

Further improvements on the system can be achieved by integrating information on multiple video frames. Another possible improvement would be to make texture analysis unsupervised to increase robustness.

6. REFERENCES

[1] S. Mori, C.Y. Suen, and K: Yamamoto, "Historical Review of OCR Research and Development," *Proc. IEEE*, vol. 80, no. 7, pp. 1029-1058, July 1992.

[2] Zhong Y., Karu K., Jain A.K., "Locating Text in Complex Color Images," *Pattern Recognition*, 28(10), pp. 1523-1535, 1995.

[3] Wu V., Manmatha R., Riseman E.M., "TextFinder: An Automatic System To Detect and Recognize Text in Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 11, November 1999.

[4] Sato T., Kanade T., Hughes E.K., Smith M.A., Satoh S., "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions," *Multimedia Systems*, 7, pp. 385-395, 1999.

[5] Ohya J., Shio A., Akamatsu S., "Recognizing Characters in Scene Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 2, February 1994.

[6] Lienhart R., Effelsberg W., "Automatic Text Segmentation and Text Recognition for Video Indexing," *Multimedia Systems*, 8, pp. 69-81, 2000.

[7] Jain A.K., Yu B., "Automatic Text Location in Images and Video Frames," *Proc. IEEE Pattern Recognition*, Vol. 31, No. 12, pp. 2055-2076, 1998.

[8] Jain A.K., Bhattacharjee S., "Text Segmentation using Gabor Filters for Automatic Document Processing," *Machine Vision and Applications*, 5, pp. 169-184, 1992.

[9] Li H., Doermann D., Kia O., "Automatic Text Detection and Tracking in Digital Video," *Computer Vision Lab. Technical Report*, University of Maryland, CAR-TR-900, December 1998.

[10] Dimitrova N., Agnihotri L., Dorai C., Bolle R., "MPEG-7 Videotext description scheme for superimposed text in images and video," *Signal Processing: Image Communication*, 16, pp. 137-155, 2000.

[11] Jain, R., Kasturi, R., Schunck, B.G., *Machine Vision*, TA1634.J35, 1995.

[12] Clark M., Bovild A.C., "Experiments in segmenting texture patterns using localized spatial filters," *Pattern Recognition*, 22(6), pp. 707 - 717, 1989.

[13] Jain A.K., Farrokhnia F., "Unsupervised texture segmentation using Gabor filters," *Proc. IEEE Int. Conf. Sys. Man. Cybern.*, pp. 14 - 19, Los Angeles, CA, November 1990.

[14] *BilVMS*, TÜBİTAK Bilten Video Management System, `http://vms.bilten.metu.edu.tr`

Video Adaptation for Transmission Channels by Utility Modeling

Özgür D. Önür[1,2], A. Aydın Alatan[1,2]
[1]Department of Electrical and Electronics Engineering, M.E.T.U.
[2]TÜBİTAK BİLTEN, Balgat, 06531 Ankara TURKEY

Abstract— **The satisfaction a user gets from watching a video in a resource limited device, can be formulated by Utility Theory. The resulting video adaptation is optimal in the sense that the adapted video maximizes the user satisfaction, which is modeled through subjective tests comprising of 3 independent utility components : *crispness*, *motion smoothness* and *content visibility*. These components are maximized in terms of coding parameters by obtaining a Pareto optimal set. In this manuscript, inclusion of transmission channel capacity into the subjective utility model of user satisfaction is addressed. It is proposed that using the maximum channel capacity as a restriction metric, certain members of the Pareto optimal solution set can be eliminated such that the remaining members are suitable for transmission through the given channel. Once the reduced solution set is obtained, an additional figure of merit can be used to pick a single solution from this set, depending on the application scenario.**

I. INTRODUCTION

The process of modifying a given representation of a video into another representation, in order to change the amount of resources required for transmitting, decoding and displaying video is defined as *video adaptation* [1]. The first reference to *utility theory* in the context of video adaptation appears in [2]. In a more theoretical approach, only a conceptual framework that models adaptation, as well as resource, utility and the relationships in between, are presented [3]. A content-based utility function predictor is also proposed [4], in which the system extracts compressed domain features in real time and uses content-based pattern classification and regression to obtain a prediction to the utility function. However, the utility value, corresponding to a given adaptation of a video, is presented as a function of the video bit-rate [4], which contradicts the subjective nature of the utility concept.

In [10], a novel method to determine an optimal video adaptation scheme, given the properties of an end-terminal, on which the video is to be displayed, is proposed. *Utility Theory* [5] is utilized to model a strictly subjective quantity, *satisfaction*, a user will get from watching a certain video clip. The satisfaction is formulated as comprising 3 independent utilities, each depending on certain video coding parameters.

In this manuscript, the effect of transmission channel capacity on the previously proposed subjective models [10] is addressed. The incorporation of the channel capacity effect to the models in [10] results in a complete formulation of the overall user experience in a multimedia delivery scenario from the content server to the mobile user terminal, as depicted in Figure 1.

II. PROPOSED ADAPTATION SYSTEM

The main aim of this paper is quantitatively determining the "satisfaction" a user gets from watching a video clip on a resource limited device, as a function of video coding parameters, the terminal device properties and also the capacity of the communication channel. Initially, the subjective satisfaction of the user is modeled ignoring channel capacity [10]. Subsequently, the effects of finite channel capacity are incorporated into the proposed model.

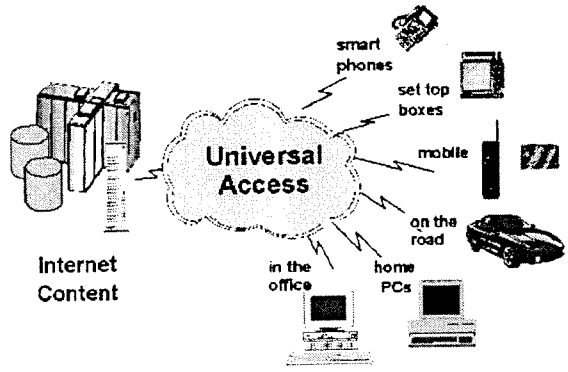


Figure 1. Delivery of the multimedia content from the media server to mobile terminal

A novel approach to obtain the utility function for the problem above, is proposed. The problem is considered as a multiple objective utility formulation. The overall utility function is decomposed into 3 independent components, such that the satisfaction associated with any one of these terms can be considered as independent from every other component. These terms are determined as:

- "Crispness" utility of a video clip,
- "Motion-smoothness" utility of a video clip,
- "Content visibility" utility of a video clip.

The reason for such decomposition is due to the perceptual independence of the proposed sub-objectives. In other words, video frames with very low distortion might be displayed in a non-smooth manner in time or a motion smooth video can independently have a very low spatial resolution. When the above decomposition is performed, the sub-objectives can be easily modeled as simpler functions of the video coding parameters by using the parametric approach of the Utility Theory [5].

A. Crispness Utility

Crispness, whose subjective nature enables it to be modeled by utility theory, is basically the perceptual similarity between the intensity edges in a digitized and compressed video, and the edges in a real-life scene, as perceived by a human viewer. The most dominating parameter, affecting the crispness of a video, is the number of bits per pixel (*bpp*) for a fixed encoder performance. In order to express the encoded *bpp* in terms of the coding parameters, the bit-rate needs to be normalized by both frame rate and spatial resolution. Hence, the first component of the overall utility function can be formulated as

$$U_{crisp}(\text{coded bits per pixel}) = U_{crisp}(CBR(CFR\,CSR)) = U_{crisp}(CBR,CSR,CFR) \quad (1)$$

where *CBR*, *CSR*, *CFR* stand for Coded Bit Rate, Coded Spatial Resolution, and Coded Frame Rate, respectively. It should be noted that all video coding parameters are referred as *coded* parameters, since these factors can be viewed differently, when video is rendered on the screen of a resource limited device.

It has been shown that perceived crispness of a video increases substantially, as *bpp* value is increased [6]. However, this increase

reaches to saturation after a range of values for *bpp* is exceeded. This saturation is due to the inability of the HVS to discern the difference in crispness of a picture, resulting from increasing *bpp* value beyond a certain point [7]. The crispness utility is also expected to depend on the *CSR* of the video, since perception of crispness for a given picture is related also to its resolution. In the light of the above observations it can be asserted that, the utility of crispness curve should have an exponential form, as expressed by the following formula where $c_1(CSR)$ is to be determined from subjective experiments. :

$$U_{crisp}(CBR, CSR, CFR) = 1 - e^{-c_1(CSR)\frac{CBR}{CFR\ CSR}} \quad (2)$$

It has been shown that the perception of crispness depends on the texture content of the image under evaluation [6]. This effect can be incorporated into the proposed model by using any metric extracted from the video describing its texture and using a modified c_1 expression which is a function of this metric. Subjective tests related to the perception of crispness have been performed for different videos, having significantly varying levels of texture and the results are presented [10]. A subset of all these subjective test results can be found in Section II.D.

B. Motion Smoothness Utility

Motion-smoothness is also another subjective phenomenon, indicating the perceptual similarity of temporal motion of an event in real world, and the motion observed through the succession of video frames. The motion smoothness of a video clip can be modeled as a function of *CFR* only, if the resource constraints of the user terminals are not taken into consideration. However, the observed frame rate during playback in a user terminal will generally not be equal to the *CFR*, due to resource limitations.

Intuitively, the frame rate, at which the "observed frame rate" deviates from the original coded frame rate, should depend on the *CBR* and *CPU*. Hence, it can be stated that the motion smoothness of a video, being observed on a user terminal, should depend on the frame rate at which the video was originally coded, the bit rate of video, and obviously, *CPU* of the end terminal. Thus, the functional representation for the second component of the utility function is determined as

$$U_{smooth}(CFR, CBR, CPU)$$

Intuitively, the motion smoothness utility is expected to increase up to a point in an exponential form with increasing *CFR* and then reach to saturation (similar to the increase in crispness utility with increasing *bpp*). This effect has been demonstrated through subjective tests for different content types [8]. The point at which the utility of motion smoothness starts decreasing due to resource limitations, should depend on the *CBR* of the video, as stated earlier. Hence, the motion smoothness utility can be modeled as follows: the exact location of the "turning point"; i.e. the frame rate at which the motion smoothness utility starts decreasing for a given bit-rate, should be determined as a function, *FR(CBR)*. However, the rate of such a decrease in utility should differ for devices with different *CPU* capabilities.

Without losing generality, in order to simplify the formulation, only two different *CPU* configurations are utilized, while modeling the dependence of motion smoothness utility on the *CPU* of the terminal device. In other words, terminal devices, having a *CPU* clock frequency higher than a predetermined threshold value, are considered as operating in *CPU High* mode, while the ones having a lower clock speed are considered to be operating in *CPU Low* mode. The corresponding utility function for *CPU Low* is expected decrease more rapidly, compared to that of *CPU High* in the *CFR>FR(CBR)* region. Based on the reasoning above, the utility function model in (3) is proposed. Note that, "time constants" of the exponential terms, sm_{0L}, sm_{1L} sm_{0H} and sm_{1H}, for *CPU High* and *CPU Low* cases are different functions of *CBR*, yielding different increase and decrease rates at each *CBR*. It has been shown that the perceived motion jerkiness of a video depends also on the viewed content [8]. The results of the subjective tests in Section II.D indicate that for a low-motion content video, motion smoothness does not decrease, i.e. enter saturation, even at the highest bit-rate.

$$U_{smooth}(CFR, CBR, CPU) = \begin{cases} 1 - e^{-sm_{0L} CFR} & , CFR <= FR_L(CBR) \\ sm_L e^{-sm_{1L}(CFR - FR_L(CBR))} & , CFR > FR_L(CBR) \end{cases} \} \text{CPU Low}$$

$$U_{smooth}(CFR, CBR, CPU) = \begin{cases} 1 - e^{-sm_{0H} CFR} & , CFR <= FR_H(CBR) \\ sm_H + 1 - e^{-sm_{1H}(CFR - FR_H(CBR))} & , CFR > FR_H(CBR) \end{cases} \} \text{CPU High} \quad (3)$$

$$FR \propto \frac{1}{CBR} \qquad sm_{L,H} = 1 - e^{-sm_{0L,H} FR_{L,H}}$$

This effect can be easily incorporated into the model by considering a video motion activity measure (e.g. MPEG-7 motion activity descriptor) and using modified *FR* and *sm* expressions that are also functions of this measure, as well as the *CBR*.

C. Content Visibility Utility

Content visibility utility is simply related to the comprehensibility and visibility of the video content with respect to its resolution and the screen size of the terminal.

The utility of the content visibility of a video clip should depend on two factors: Initial *CSR* of the video and the *screen size* of the user terminal. A video, can only be viewed partially, i.e. either cropped or down sampled, on a terminal whose screen size is smaller than the *CSR* of this video. The results of the subjective tests in the proceeding section show that cropping results in reduced user satisfaction, as expected. On the other hand, down sampling does not further reduce the satisfaction and should yield a saturated satisfaction after *CSR* exceeds the screen size. For both of these cases, the final component of the utility function is prototyped as follows:

$$U_{cv}(CSR, Screen\ Size)$$

Considering only the cropped case, the utility of the content visibility of a video clip is expected to increase in a similar fashion to (2) and (3), up to the point at which the spatial resolution becomes equal to the screen size of the terminal. After that point, in case of cropping, the utility is expected to decline conforming to the following equation:

$$U_{cv}(CSR, ScreenSize) = \begin{cases} 1 - e^{-s_1 CSR} & CSR \le ScreenSize \\ s e^{-s_2 (CSR - ScreenSize)} & CSR > ScreenSize \end{cases} \quad (4)$$

$$s = 1 - e^{-s_1 ScreenSize} \qquad s_1 \propto \frac{1}{ScreenSize} \qquad s_2 \propto \frac{1}{ScreenSize}$$

The parameters s_1 and s_2 should be both inversely proportional with the screen size of the terminal, since the increase or decrease in utility is expected to happen more abruptly in devices with smaller screen sizes. Similar to prior discussions, it should be noted that different types of content type might affect the proposed models in various ways. For instance, on close-up shots where the content fills the whole screen, the utility might decrease more suddenly, when the video frame is cropped, whereas on shots for which the scene mainly consists of a repeating pattern, such an effect may not be the case. These effects can be incorporated into the model by using a metric that defines the distribution of the content within the scene and using modified s_1, s_2 expressions which are functions of this metric.

D. Subjective Tests for Determining Utility Functions

In the next step, the utility associated with each sub-objective is

determined for various video coding parameters and terminal characteristics by a series of subjective evaluation experiments. These experiments are performed in accordance with the principles stated in *ITU-R 500-11 Subjective Television Picture Assessment Standard* [8]. The tests were performed on a *Siemens Pocket LOOX 600* Personal Digital Assistant (PDA). The selected test method is the *Double Stimulus Impairment Scale (DSIS)* [8].

Figure 2(a) illustrates the results of subjective tests related to the utility of crispness for a high texture image from a sitcom.

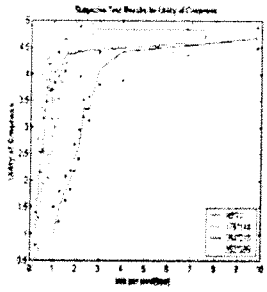


Figure 2(a): Subjective Test Results for Utility of Crispness for high textured content.

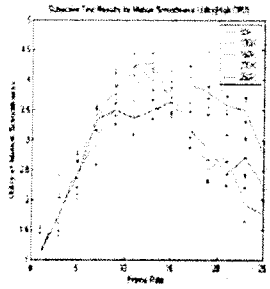


Figure 2(b): Subjective Test Results for Utility of Motion Smoothness for high motion content (High CPU)

Figures 2(b), 3(a), and 3(b) illustrate the results of subjective tests related to the utility of motion smoothness. While Figures 2(b) and 3(a) demonstrate the results for a high-motion soccer video for the high CPU and the low CPU cases respectively, Figure 3(b) shows the results for a very low motion content video, consisting of an anchorman with only limited head motion for the high CPU case. Figure 3(b) shows that even for a video encoded with 250 Kbits/s, the end terminal is able to decode this low-motion video in real time. This result is expected, since motion compensation, which is an computationally expensive phase of the decoding process (quite demanding especially for a PDA), is less utilized for such a static video, in comparison to the active video of Figures 2(b) and 3(a). Figure 4(a) illustrates the results of subjective tests related to the utility of content visibility for a sequence containing a close-up recording of a dialogue scene. Figure 4(b) shows the system recommended videos, being displayed on a typical PDA for which the simulations are performed. These images are captured from video sequences, available at `www.eee.metu.edu.tr/~alatan/adapt.`

In order to obtain the overall utility equation, it is necessary to determine the parametric functions, utilized in the proposed models for the individual utilities. Using the results of the subjective evaluation tests, these functions are obtained in terms of *CBR*, *CFR* and *CSR* by simple least squares fitting.

E. Finding Optimal Set of Encoding Parameters

After determining all the utility components, the next goal is to determine the set of encoding parameters which maximize these components for a given device. For such multiple criteria optimization problems, finding the Pareto optimal solution set is often the first step towards obtaining the optimal solution, since a *dominated solution* can not be optimal [10].

In order to determine the Pareto optimal set, a 3-D parameter space, formed from bit-rate (BR), frame rate (FR) and the spatial resolution (SR), can be sampled, so that a finite set of points (BR,FR,SR) is obtained. The values for the individual sub objectives are calculated for each point in this space for a given user terminal. Note that at this stage, each (BR,FR,SR) triplet together with the user terminal parameters is mapped into another vector $U(U_C,U_{MS},U_{CV})$, composed of the utility values for the individual sub utilities. In order to determine the optimal solution, the non-dominated U vectors are selected form the Pareto optimal set of utility components [11]. This set, being Pareto optimal, contains only the vectors for which it is not possible to find another solution vector having *all* the component utilities larger than the corresponding component utilities of the member vector. The Pareto optimal set can be further refined, by discarding the solutions for which the value of any one of the component utilities is so low that the dissatisfaction associated with it impairs the judgment of the overall utility. In other words, any of the individual utilities of a member vector can not be less than a predetermined threshold (which is heuristically chosen as 20% of the maximum possible utility). Such a restriction reinforces the assumption of independence between the component utilities, since it does not allow severely impaired videos to enter the solution set.

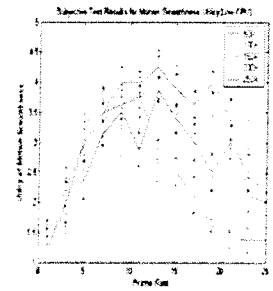


Figure 3(a): Subjective Test Results for Utility of Motion Smoothness for high motion content (Low CPU)

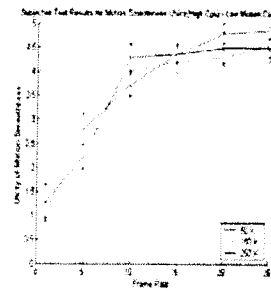


Figure 3(b): Subjective Test Results for Utility of Motion Smoothness for low motion content (High CPU)

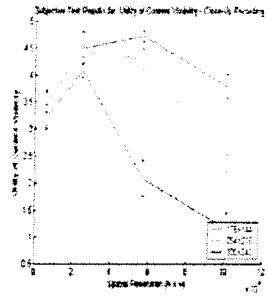


Figure 4(a): Subjective Test Results for Utility of Content Visibility Close-Up Recording

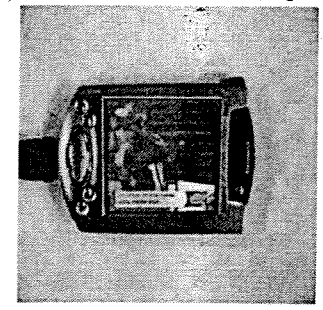


Figure 4(b): Compaq IPAQ displaying video at recommended parameters.

Once the Pareto optimal set is determined, the effect of finite channel capacity can be considered. For a given maximum channel capacity, the members of the Pareto optimal set having associated bit-rate (BR) values higher than this capacity are discarded from the Pareto optimal set. The remaining members are all suited for transmission through the given channel. In order to choose a specific solution from the remaining members, an additional figure of merit needs to be selected. In the simulations presented in the following section the solution having the highest associated bit-rate, i.e. the bit rate closest to the channel capacity, has been chosen so as to utilize the channel to the fullest extent. Choosing another criterion to select a specific member of the set, such as having the highest motion smoothness utility or highest crispness utility are equally valid.

III. SIMULATIONS

A. *Effects of Finite Channel Capacity*

3D parameter space is sampled into 6000 discrete points. Then, the Pareto optimal solution set is obtained by using the procedure outlined in Section II.E. If there are no restrictions on the channel capacity, the Pareto optimal set contains 1234 members. Figure 5 shows the members of the Pareto optimal set when the maximum channel capacity is restricted to 75 Kbits/s. 92 members have associated bit rates lower than the specified capacity. The marked solution is chosen, as it has the highest associated bit-rate, i.e. the bit rate closest to the channel capacity as specified in the previous section.

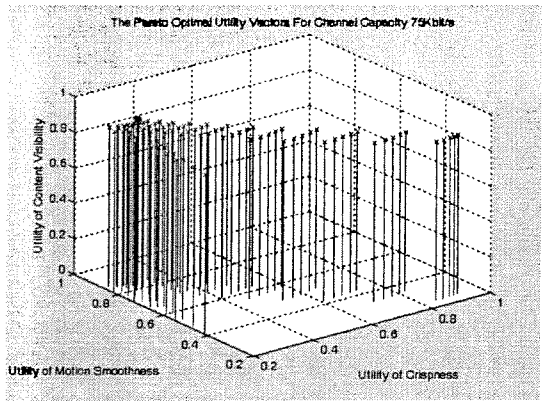


Figure 5 : The Pareto optimal utility vectors for a channel with capacity 75 Kbits/s

When the maximum channel capacity is increased to 100 Kbits/s, 211 members out of the initial 1234 remain in the solution set. The solution that has the highest associated bit-rate is once again marked in the plot. The total execution time required to obtain the Pareto optimal set is slightly less than 2 seconds in a Intel Pentium III laptop computer.

IV. CONCLUSIONS

The main contribution of this paper is inclusion of transmission channel capacity into the previously proposed subjective utility models for user viewing satisfaction in resource limited devices. It has been shown that using the maximum channel capacity as a restriction metric, certain members of the Pareto optimal solution set can eliminated such that the remaining members are suitable for transmission through the given channel. Once the reduced Pareto optimal set is obtained, an additional figure of merit can be used to pick a single solution from this set depending the application scenario. Finally, it should be stated the proposed utility models for video adaptation scenario is quite generic in the sense that different content types or channel capacity restrictions, can still be incorporated into these models.

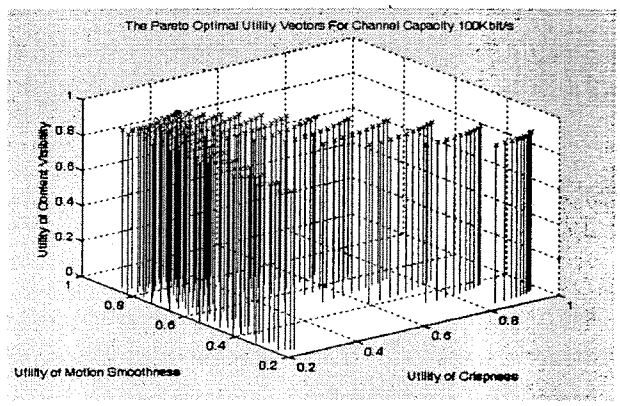


Figure 6 : The Pareto optimal utility vectors for a channel with capacity 100 Kbits/s

V. REFERENCES

[1] Y. Neuvo , J. Yrjanainen, "Wireless Meets Multimedia – New Products and Services," Proc. of IEEE ICIP 2002

[2] P. Bocheck, Y. Nakajima and S.-F. Chang, "Real-time Estimation of Subjective Utility Functions for MPEG-4 Video Objects," Proceedings of IEEE Packet Video Workshop (PV'99), New York, USA, April, 1999.

[3] S. F. Chang, "Optimal Video Adaptation and Skimming Using a Utility-Based Framework," Tyrrhenian International Workshop on Digital Communications Capri Island, Italy September 2002

[4] Y. Wang, J.-G. Kim, S.-F. Chang, "Content-Based Utility Function Prediction For Real-Time MPEG-4 Video Transcoding," Proc. of IEEE ICIP 2003, Barcelona SPAIN.

[5] A. L. Golub, "Decision Analysis: An Integrated Approach," John Wiley and Sons Inc, 1997

[6] M. Ardito, M. Barbero, M. Stroppiana, M. Visca, "Compression and Quality," Proceedings of the International Workshop on HDTV, 1994

[7] K. Boff, L. Kaufman, & J. Thomas (Ed.), "Handbook of Perception and Human Performance : Sensory Processes and Perception," John Wiley and Sons Inc, 1986.

[8] Ronnie T. Apteker, James A. Fisher, Valentin S. Kisimov, and Hanoch Neishlos, "Video acceptability and frame rate," IEEE Multimedia Volume: 2 , Issue: 3 , Fall 1995 Pages:32 – 40

[9] Recommendation ITU-R BT.500-11 "Methodology for the Subjective Assessment of the Quality of TV Pictures," 2002.

[10] Ö. D. Önür, A. A. Alatan, "Optimal Video Adaptation for Resource Constrained Mobile Devices Based on Utility Theory," WIAMIS 2004, Portugal.

[11] Ralph E. Steuer, "Multiple Criteria Optimization: Theory, Computation and Application," John Wiley and Sons Inc, 1985

Optimal Video Adaptation for Resource Constrained Mobile Devices Based on Utility Theory

Ö. D. Önür[1,2], A. A. Alatan[1,2]

[1]Department of Electrical and Electronics Engineering, M.E.T.U.
[2]TÜBİTAK BİLTEN

ABSTRACT

The diversity of user terminals that can be used to access multimedia content, necessitates the tailoring of the content according to the computational capabilities of the terminals. Optimal video adaptation based on multimedia playback device characteristics is addressed. The adaptation is optimal in the sense that, the adapted video maximizes the user satisfaction. Utility Theory is used to formulate the satisfaction a user gets from watching a video. A novel approach to divide the user satisfaction into three independent components is proposed. The individual components are modelled by exponential curves and their weighted sum is used as the overall satisfaction or 'utility' function. The combination of trade-off weights that result in highest user satisfaction are obtained through experiments with different device characteristics. The global maximum of the utility function is found by simulated annealing.

1. INTRODUCTION

One of the major challenges obscuring the path towards enjoying mobile multimedia, is delivering the multimedia content to various mobile terminals across a wide range of networks[1].This issue has attracted considerable attention in the signal processing community. The concept of Universal Multimedia Access (UMA) has been devised to deal with this problem. UMA aims to deliver rich multimedia content to terminals with different processing capabilities, across heterogeneous networks, tailored to specific user preferences. This task is quite challenging. The solution requires frequently adjusting the resource requirements of the video as the video traverses the interconnected networks from the satellite to the mobile terminal, because different networks and terminals will certainly have different capacities and characteristics. The process of changing a given representation of a video to a different representation, in order to change the amount of resources required to transmit, decode and playback the video is called *video adaptation.*

In this paper, a method to compute the optimal video adaptation scheme, given the properties of the end terminal on which the video will be displayed, is proposed. *Utility Theory* is utilized to construct models, that are fitted to results of subjective evaluation tests, to formulate the satisfaction a user gets from watching a certain video clip. The *'utility function'* obtained as described above is maximized to obtain the representation of the video that results in highest user satisfaction.

In the next section the related work in the literature is investigated. In Section 3, utility theory is introduced. In Section 4, the proposed system is detailed. Simulations are presented in Section 5 and the paper ends with conclusions in Section 6.

2. RELATED WORK

In [3] a conceptual framework to model adaptation, resource, utility and the relationships between them is presented. The term entity is defined to be a chunk of video that shares a certain consistent property, and it is also labelled as the data unit that undergoes adaptation. Objective measures such as the SNR, coherence, temporal smoothness are used to measure utility. The optimal video adaptation problem is formulated as, finding the adaptation operation that maximizes the utility of the adapted entity given the original entity and resource constraints. Objective measures fail to model human satisfaction adequately, to obtain an acceptably accurate model, a multitude of attributes need to be extracted from the video, and this significantly increases the computational complexity of the system.In [2] a system to adapt multimedia web content to match the capabilities of the device requesting it, is introduced. The system has two components; The InfoPyramid which creates and stores, multimodal and multi-resolution representations of the multimedia content and the Customizer that selects the representation of the content from the various available versions in the InfoPyramid. Considering the diversity of the terminals that can be used to access multimedia content, an optimal representation for each one of the different terminals can not be obtained from a predetermined set of representations.

3. UTILITY THEORY

The fundamental motive of utility theory is, to represent the satisfaction or *expected utility* of a resource as a function of the amount of that resource. If this can be accomplished, the most efficent allocation of a resource can be identified among various alternatives. There are two methods to obtain the utility function of a resource in utility theory. Both of these methods rely on subjective utilities provided by individual(s) representing the community for which the utilities need to be determined :

1. Eliciting the utility values directly from the individual, by presenting the best and worst possible results and asking the individual to determine the relative satisfaction of all the remaining points of the utility function.
2. Assigning an exponential form to the utility function such as

$$U(x)=x^c \quad \text{OR} \quad U(x)=(1-e^{-x/c}) \qquad (3.1)$$

and trying to estimate the value of the parameter 'c' in the above equations, again by presenting the best and worst possible results and asking the evaluator to provide a few

more points to be able to obtain a reliable estimate for the value of 'c'.

In some cases it might be necessary to consider multiple objectives when trying to find the utility associated with an alternative, In other words the total satisfaction might depend on more than one kind of resource. In such a case, if the satisfaction on any one of the objectives is independent from the satisfaction from every other objective, the additive utility function can be used to obtain the total satisfaction as illustrated in Equation (3.2). The additive utility function is simply the weighted sum of the individual objectives. The weights associated with the components of the overall utility function in (3.2), represent the relative importance of the individual objectives.

$$\begin{bmatrix} Total \\ Satisfaction \end{bmatrix} = w_1 U(obj_1) + w_2 U(obj_2) + \ldots + w_n U(obj_n) \qquad (3.2)$$

4. PROPOSED SYSTEM

The main aim of this paper is to accurately determine the "satisfaction" a user gets from watching a video clip on a resource limited device, as a function of video coding parameters and the terminal device properties. For a given terminal (CPU and the Screen Size values are taken into consideration), the user satisfaction is evaluated as the video coding parameters i.e. bit rate, frame rate and spatial resolution are varied. Constructing a utility function for this problem requires conducting a very large number of experiments even if the second method illustrated by Eqn. (3.1) is used. Since the utility is a function of 5 different variables, expressing the utility as a simple exponential function as in (3.1) is actually not possible.

A novel approach to obtain the utility function for the above problem is proposed. The problem is considered as a multiple objective utility formulation. The overall utility function is decomposed into three independent components such that, the satisfaction associated with any one of the components is independent from every other component. These components are determined as:

1. the "crispness" utility of a video clip
2. the "motion-smoothness" utility of a video clip
3. the spatial resolution utility of a video clip

The reason of such a decomposition is the independence between the components, and the fact that these com-ponents can be expressed as simple functions of the video coding parameters previously mentioned. The components will be referred to as sub-objectives from this point on.

4.1 Crispness

The most accurate measure of the crispness of a video is probably the number of bits encoded per pixel(bpp). In order to express encoded bpp in terms of the coding parameters, the bit-rate needs to be divided by both frame rate and spatial resolution. In other words, the first component of the overall utility function should depend on all three video coding parameters. Hence, the first component of the overall utility function, the utility or satisfaction a user will get from the crispness of a video clip, can be formulated as

$$\begin{aligned} U_{crisp}(CBR,CSR,CFR) &= U_{crisp}(CBR/(CFR\ CSR)) \\ &= Ucrisp(coded\ bits\ per\ pixel) \end{aligned} \qquad (4.1)$$

Where *CBR* stands for Coded Bit Rate, *CSR* stands for Coded Spatial Resolution, and *CFR* stands for Coded Frame Rate. It should be noted that all the video coding parameters are referred as 'coded' parameters. The reason for such a referral is that when the video is rendered on a client device, the video parameters of the video being played back, may not be exactly the same as the parameters that were originally coded. The reason for this inconsistency is due to the resource constraints of the user terminals. The phrase "coded" is used to emphasize that the values being used here, are the original encoding values of the parameters, forced at the encoder.

4.2 Motion Smoothness

The motion smoothness of a video clip, is characterized by the coded frame rate. The frame rate at which the motion loses smoothness or in other words the frame rate at which, the observed frame rate deviates from the original coded frame rate depends on the encoded video bit-rate. This is expected, since decoding a high bit rate video requires significant computational resources, and after a certain bit rate is exceeded the CPU won't be able to decode the video in real time. In light of the above discussion, it can be said that the motion smoothness of a video that will be observed on a user terminal, should depend on the frame rate at which the video was originally coded, the bit rate of the video, and CPU of the end terminal. Thus, the second component of the utility function is determined as

$$U_{smooth}(CFR,CBR,CPU)$$

4.3 Spatial Resolution

Intuitively, the utility of the spatial resolution of a video clip should depend on two factors: Spatial resolution that the video is initially coded with and the *screen size* of the user terminal on which the video is to be displayed. One can easily realize that if a video, having a spatial resolution larger than the screen size of the terminal, is transmitted to the terminal, a portion of the video must be clipped in order to display the video on that device. This will inevitably result in reduced user satisfaction and should be avoided, if possible. The final component of the utility function is prototyped as follows:

$$U_{size}(CSR.ScreenSize)$$

4.4 Utility Function Generation

The satisfaction for each of the sub-objectives mentioned in the previous section is assumed to be independent of the satisfaction on every other sub-objective. For example, the satisfaction a user gets from the motion smoothness of a video has no relation with the crispness of the same video. Therefore, one can use the *additive utility function* [5] to determine the total satisfaction a user will get from watching a certain video. Thus the final equation for the utility can be obtained as follows:

$$\begin{aligned} U = {} & w_1 U_{crisp}(CBR, CSR, C FR) + w_2 U_{smooth}(CFR, CBR, CPU) \\ & + w_3 U_{size}(CSR, Screen\ Size) \end{aligned} \qquad (4.2)$$

The weights, w_1,w_2 and w_3 associated with the terms of the utility function, are to be determined by simulations. The first step, in the process for obtaining the utility function, is determining the general forms of the component curves, for each sub-objective.

4.4.1 Crispness Term in Utility Function

The satisfaction a user gets from crispness of a video, should increase substantially as the bpp value is increased from zero, but this increase is expected to reach saturation

after a certain value of the bpp. This saturation is due to the inability of the HVS to discern the difference in crispness of a picture resulting from increasing the bpp value beyond a certain point. Hence, increasing bpp value further is not expected to result in a substantial increase in user satisfaction on crispness.

In the light of the above observations and postulates of utility theory[5], it can be asserted that, the utility of crispness curve, should have an exponential form as expressed by the following formula

$$U_{crisp}(CBR,CSR,CFR)=1-e^{-c_1\frac{CBR}{CFR\,CSR}} \qquad c_1 \alpha\ CSR \qquad (4.3)$$

The reason for the inclusion of the expression c_1, and its relation with CSR, is as follows; It should be noted that the bpp value obtained by the relation (CBR/CFR*CSR) in the above formula, is the bpp value after compression. Since compression algorithms use redundancy in a picture to compress data[4], they have higher compression rates, while compressing pictures with larger spatial resolution, if pictures having more or less the same frequency content are considered. In this case, when pictures, differing only in their spatial resolution are compared, the picture with the larger spatial resolution is compressed much more efficiently, since it contains greater redundancy compared the one with the lower spatial resolution. Therefore, bpp value required to code a picture with a given crispness, is smaller for the pictures having higher spatial resolutions. In order to account for this fact, a function c_1 has been included in the above formulation. Larger the value of c_1, higher the crispness of a video for a given bpp value. Hence, c_1 should be directly proportional with the CSR.

4.4.2 *Motion Smoothness Term in Utility Function*

While increasing CFR, the second component of the utility function, utility of motion smoothness, should increase up to a point, in an exponential expression, similar in form, to the utility of crispness, as shown in Eq. (4.4). The point at which the utility of motion smoothness starts decreasing, due to resource limitations, should depend on the coded bit rate of the video, as stated earlier. In the formulation above, the function FR(CBR) determines the exact location of this point. i.e. for a given bitrate, if the frame rate is higher than FR(CBR) the motion smoothness starts decreasing. The formulation of the dependence of motion smoothness utility, on the clock-speed of the CPU of the terminal device, is simplified.

$$U_{smooth}(CFR,CBR,CPU)=\begin{cases}\begin{cases}1-e^{-a_0 CFR} & CFR\le FR_L(CBR)\\ a_1 e^{-c_2(CFR-FR_L(CBR))} & CFR>FR_L(CBR)\end{cases} & CPU\ Low\\ \begin{cases}1-e^{-a_0 CFR} & CFR\le FR_H(CBR)\\ a_1+1-e^{-c_3(CFR-FR_H(CBR))} & CFR>FR_H(CBR)\end{cases} & CPU\ High\end{cases} \qquad (4.4)$$

$$FR\ \alpha\ 1/CBR$$

$$a_1=1-e^{-a_0 FR}$$

It is performed for two different clock speeds only, one representing *highCPU*, and the other representing *lowCPU*. Any CPU value input to the system, will be classified as high or low CPU according to a simple threshold, to make the analysis manageable.For frame rates, where the utility is increasing (up to the limit defined by FR), the utilities of the high CPU and the low CPU cases are assumed to be the same. This is reasonable, since the observed and the coded frame rates are the same up to that point, and a particular frame rate, gives the same utility across all platforms unless distorted by the resource constraints. The a_0 term in Eq. (4.4) is a constant that will be determined based on the results of the performed subjective tests.FR should be inversely proportional to CBR, i.e. as the bit rate is increased, the observed frame rate starts deviating from the coded frame rate, at smaller frame rates. The function FR is different for the cases of high CPU and low CPU. Severe degardation in motion smoothness utility is expected as the CBR value increases beyond the limits of the capacity of the CPU. To account for this fact, two functions c_2 and c_3 are used in the above formulation. Notice that in both expressions of Eq. (4.4), for larger c_2 or c_3 the utility drops faster, so using c_2 and c_3 in direct proportionality to CBR, the desired form for the utility curves can be obtained. The point a_1 in both expressions is the value of utility, at which the functions start decreasing.

4.4.3 *Spatial Resolution Term in Utility Function*

Finally, the utility of the spatial resolution of a video clip is expected to increase in a similar fashion to Eqns. (4.3) and (4.4), up to the point where the spatial resolution, becomes equal to the physical screen size of the user terminal. After that point, the utility is expected to decline conforming to the following equation:

$$U_{size}(CSR,ScreenSize)=\begin{cases}1-e^{-a_{21}CSR} & CSR\le ScreenSize\\ c_4 e^{-a_{22}(CSR-ScreenSize)} & CSR>ScreenSize\end{cases}$$

$$c_4=1-e^{-a_{21}ScreenSize} \qquad (4.5)$$

$$a_{21}\ \alpha\ 1/ScreenSize$$

$$a_{22}\ \alpha\ 1/ScreenSize$$

The functions a_{21} and a_{22} are used to account for this fact. Both are inversely proportional with the screen size of the terminal. Note that, larger a_{21} leads to a steeper increase, in the increasing portion of the utility function, and larger a_{22} means a steeper decrease in the declining portion of the utility. Since the increase and decrease in utility is expected to change more abruptly in devices with smaller screen sizes, the inverse proportionality of a_{21} and a_{22} with screen size is reasonable.

4.5 Subjective Test for Utility Function

The next step towards determining the components of the utility function uniquely is a series of subjective evaluation experiments. The experiments are performed separately for each component. While an experiment on one of the components is being performed, the video coding parameters not affecting the utility of that component are kept constant. The steps followed while performing the experiments are as follows:

The evaluators are first shown the videos that are considered the best and the worst for the particular component of the utility function, for which the experiment is being performed. For example, the video having a frame rate of 1 fps is shown as the worst sample for the utility of the motion smoothness component, whereas the video having a frame rate of 25 fps, is presented as the best sample. While the utility of the worst sample is fixed at 0, and the utility of the best sample is fixed at 100, as required by utility theory. Obviously, the spatial resolution of the sample videos is held constant (at 176x144 pixels) throughout the test performed for motion smoothness. The evaluators are shown videos, coded with different values of the parameter(s) that has an influence on the component of the utility being tested. The evaluators are asked grade those samples with grades ranging between 0 and 100, according to the satisfaction they get from watching that video. The important point here is that they are asked to evaluate the videos, only according to the component being

tested. For example, if the motion smoothness was being tested they are asked to express their satisfaction only regarding the motion smoothness of the video, ignoring their satisfaction regarding the crispness or the size of the video.

4.6 Fitting Parameters of Utility Function

In order to illustrate the procedure, the determination of the expression c_1 used in Eqn. (4.3) will be explained, the other expressions and constants are determined in a similar manner. First by rearranging Eqn. (4.3),

$$\frac{CFR\,CSR}{CBR}\ln(1-\tilde{U}_{comp})+c_1=0 \qquad (4.6)$$

Eqn. (4.6) should hold for all test data, if the Eqn. (4.3) has no errors. Minimizing sum of squared errors with respect to unknown c_1 gives the optimal c_1 value.

$$\sum_{\substack{\text{experimental}\\ \text{utility values}}}\left(\frac{CFR\,CSR}{CBR}\ln(1-\tilde{U}_{comp})+c_1\right)^2 \qquad (4.7)$$

In order to find c_1, which makes Eqn. (4.7) minimum; first one should take its derivative and then equate to zero before solving for c_1.

$$c_1=\frac{1}{n}\sum_{\substack{\text{experimental}\\ \text{utility values}}}\frac{CFR\,CSR}{CBR}\ln\left(\frac{1}{1-\tilde{U}_{comp}}\right) \qquad (4.8)$$

In curve fitting phase, MATLAB is utilized for its *least squares fitting algorithm*, to fit curves to the observed data. The method minimizes *summed square of residuals* between the observed data and the fitted curve. The summed square of residuals is given by the following formula

$$S=\sum_{i=1}^{n}r_i^2=\sum_{i=1}^{n}(y_i-y_i')^2 \qquad (4.9)$$

All the curves, which are fitted to the functions, are selected as exponential fits, with the exception of a_{21} of Eqn. (4.5), since the exponential fit yields a high residue for that case. Finally, for maximizing the obtained utility function, a stochastic optimization technique, known as *simulated annealing* is used. The advantage of stochastic methods over their deterministic counterparts is having a lower probability of getting stuck in local maxima/minima [6,7,8].

5. SIMULATIONS

When the unknown expressions in Equations (4.3)-(4.5) are fitted exponential functions in accordance with the procedure presented in Section 4.6, it is observed that the model proposed in Section (4.4) accurately accounts for the observed experimental data. The final stage in maximizing the utility function is obtaining the values of the trade-off weights, w_1, w2, w3, used in Eqn. (4.2), at which the utility function has the maximum value. At this point, the question is whether the values of trade-off weights, at which the utility function is maximum, change for different values of screen size and CPU; i.e. for different user terminals. A series of simulated annealing experiments are performed to answer this question by obtaining the optimum values of the trade-off weights, for four different user terminals having CPU clock speed and screen size values of (400MHz, 176x144), (400MHz, 352x288), (200 MHz, 176x144) and (200MHz, 352x288). The trade off weights are varied between 0.1 and 0.5, in 6 discrete levels (0.1,0.2,0.3,0.33,0.4,0.5) in such a way that, the sum ($w_1 + w_2 + w_3$) always amounted to unity. It is seen that the value of the weights that maximize the utility function do not change for different user terminals. Although the terminals, on which the experiments, are performed do not span the entire range of terminals, the results are expected to be approximately the same. It is observed that for the four different user terminals described above, the values of the trade-off weights that maximize the utility function are W_1=0.1, W_2=0.4, W_3=0.5.At this point, the utility function is uniquely determined. The optimal values of the video encoding parameters, CBR,CFR and CSR, can be found for any given terminal device. The system needs only the CPU and the screen size of the terminal device to compute the optimal values of the video coding parameters, using the utility function with the weights, as determined above. A user can also specify different values for the weights according to his/her personal preferences.

6. CONCLUSIONS

The main contribution of this paper is the decomposition of the satisfaction a user gets from watching a video into three conceptually independent components, as the satisfaction resulting from the crispness of a video, the satisfaction resulting from the motion smoothness of a video and the satisfaction resulting from the spatial resolution of a video. It has been observed that such decomposition enables, more accurate subjective evaluation of the user satisfaction. This in turn makes possible, precise modelling of the user satisfaction in terms of the video coding parameters. In summary, a novel methodology, for accurately modelling user satisfaction, using utility theory is proposed and implemented.

The system implemented in the paper work successfully determines the representation of a video that will result in maximum user satisfaction, for a specific user terminal with given CPU and screen size.

7. REFERENCES

[1] Y. Neuvo , J. Yrjanainen, "Wireless Meets Multimedia – New Products and Services," IEEE ICIP 2002

[2] R. Mohan, John R. Smith, Chung-Sheng Li, "Adapting Multimedia Internet Content for Universal Access ," IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 1, NO. 1, MARCH 1999

[3] S. F. Chang, "Optimal Video Adaptation and Skimming Using a Utility-Based Framework," Tyrrhenian International Workshop on Digital Communications (IWDC-2002) Capri Island, Italy September 2002

[4] B. G. Haskell, "Digital Video: An Introduction to MPEG-2, " Kluwer Academic Publishing, 2000

[5] A. L. Golub, "Decision Analysis : An Integrated Approach," John Wiley and Sons Inc, 1997

[6] R. O. Duda, P. E. Hart, D. G. Stork, "Pattern Classification," Wiley Interscience, 2001

[7] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, "Optimization by Simulated Annealing," Science 13 MAY 1983, Volume 220, Number 4598

[8] A. M. Tekalp, "Digital Video Processing," Prentice Hall Signal Processing Series,1995

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Networks 48 (2005) 489–501
Computer Networks
www.elsevier.com/locate/comnet

A simple and effective mechanism for stored video streaming with TCP transport and server-side adaptive frame discard

Eren Gürses [a], Gozde Bozdagi Akar [a,*], Nail Akar [b]

[a] *Department of Electrical and Electronics Engineering, Middle East Technical University, 06533 Ankara, Turkey*
[b] *Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey*

Received 5 October 2003; received in revised form 2 July 2004; accepted 24 October 2004
Available online 30 December 2004

Responsible Editor: B. Baykal

Abstract

Transmission control protocol (TCP) with its well-established congestion control mechanism is the prevailing transport layer protocol for non-real time data in current Internet Protocol (IP) networks. It would be desirable to transmit any type of multimedia data using TCP in order to take advantage of the extensive operational experience behind TCP in the Internet. However, some features of TCP including retransmissions and variations in throughput and delay, although not catastrophic for non-real time data, may result in inefficiencies for video streaming applications. In this paper, we propose an architecture which consists of an input buffer at the server side, coupled with the congestion control mechanism of TCP at the transport layer, for efficiently streaming stored video in the best-effort Internet. The proposed buffer management scheme selectively discards low priority frames from its head-end, which otherwise would jeopardize the successful playout of high priority frames. Moreover, the proposed discarding policy is adaptive to changes in the bandwidth available to the video stream.

Keywords: Video streaming; Congestion control; Adaptive frame discarding; Explicit congestion notification; Differentiated services

1. Introduction

Transmission of high quality video over the Internet Protocol (IP) networks has become commonplace due to recent progresses in video compression and networking disciplines, the

* Corresponding author. Tel.: +90 312 2102341.
E-mail addresses: gurses@eee.metu.edu.tr (E. Gürses), bozdagi@eee.metu.edu.tr (G.B. Akar), akar@ee.bilkent.edu.tr (N. Akar).

1389-1286/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2004.10.015

development of efficient video coders/decoders, the increasing interest in applications such as video on demand, videophone, and video conferencing, and the ubiquity of the Internet. However, that are certain technical challenges to be overcome for efficiently transmitting video over IP networks; see for example the references [1] and [2] for an introduction to the topic. These challenges stem from the mismatch between the strict bandwidth, delay, and loss requirements of the video applications and the best-effort current Internet, which was originally designed around data applications that can tolerate loss and delay. Moreover, the instantenous bandwidth available to a certain user or application changes in all time scales because of the very dynamic nature of the Internet, making the problem even more challenging. These characteristics of the Internet led to the rise of network-adaptive video applications for providing smooth playout at the receiving client.

This paper addresses the problem of TCP-friendly on-demand streaming of temporally scalable stored video over the Internet using server-side adaptive frame discarding. In a stored video-on-demand system, the server prestores the encoded video and transmits it on demand to a client for playout in real time. The client buffers the data and starts playout after a short delay in the order of seconds (called the playout delay and denoted by T_p). We assume a fixed T_p throughout the paper as opposed to the adaptive playout schemes where the client buffering delay is varied with respect to the network conditions [3,4]. It is this tolerability to larger playout delays that distinguishes the stored video streaming problem from other video networking applications like videophony, video conferencing, and live video streaming. It is also very desirable that once the playout begins, it should be able to playout without any interruption (i.e., smooth playout) until the end of the video streaming session. Moreover, such a transmission strategy should not jeopardize the data flows on the same network path which use TCP as their transport protocol, which is referred to as the "TCP-friendliness" requirement [5–7].

For network-adaptive video transmission over IP networks, the server adapts its video injection rate into the network to the instantenous available bandwidth in the network. Several mechanisms are proposed for rate adaptation including stream switching as in the SureStream technology provided by RealSystem G2 [8,9], rate-adaptive video encoding/transcoding [1], or joint use of scalable coding (i.e., layered coding) and rate shaping via server-side selective frame discard [10]. Bitstream switching does not offer a fine granularity since there are only a few bitstreams available among which the streaming server ean switch. Rate-adaptive encoding is more appropriate for live video streaming or interactive video applications as opposed to the stored video streaming problem we discuss in this paper. In our work, we therefore focus on rate adaptation using scalable encoded bitstreams. Scalable video codecs generate two or more bit streams, one carrying the most vital video information, called the base layer (BL), and the others carrying the residual information to enhance the quality of the base layer, which is referred to as the enhancement layers (EL) [11]. If there is a single EL, then the corresponding scalable coding is called 2-layer. Several scalable video-coding techniques have been proposed over the past few years for real-time Internet applications in the form of several video compression standards such as MPEG-2/4 and H.263/H.264 [11–15]. The types of scalability which are defined in these standards can be categorized as temporal, spatial, SNR, and object (only for MPEG4) scalability; see [16] for a general overview of layered coding. In these structures, base and enhancement layers are precoded at encoding time, and therefore their rates cannot be adjusted at transmission time. Therefore, server-side selective frame discard mechanisms are proposed for rate adaptation of scalable video. These discard mechanisms intelligently decide to drop some EL frames with the goal of increasing the overall quality of the video by taking network constraints and client QoS requirements into consideration [10]. The more recent Fine Grained Scalability (FGS) coding (see [17]) in which the enhancement frame can be encoded independently with an arbitrary number of bits and the bit rate can thus be adjusted at transmission time for finer granularity is left outside the scope of the current paper. We limit the focus of this paper by using a 2-layer temporal scalability

video encoding scheme provided by H.263 version 2 (H.263+) [13] although we note that our results also apply to other 2-layer scalable video encoding schemes.

Besides network adaptivity, another challenging issue for the stored video streaming problem over the Internet is to provide inter-protocol fairness. Transmission Control Protocol (TCP) is the de-facto transport protocol for data in the current Internet. TCP is designed to offer a fully reliable service which is suitable for applications like file transfers, e-mail, etc. On the other hand, the alternative transport protocol User Datagram Protocol (UDP) used by many current streaming applications does not possess congestion control. Consequently, when UDP and TCP flows share the same link, TCP flows reduce their rates in case of a packet drop. This leaves most of the available bandwidth to unresponsive UDP flows leading to starvation of TCP traffic in case of substantial UDP load. Some believe that the current trend in using UDP as the transport layer without congestion control can lead to a congestion collapse of the Internet due to the rapid growth of such applications like Internet telephony, streaming video, and on-line games [5]. Taking into consideration the dominance of TCP in today's Internet traffic, it is therefore desirable that the throughput of a video streaming session be similar to that of a TCP flow under the same network circumstances (i.e., two sessions simulatenously using the same network path). Such a mechanism is called TCP-friendly and TCP friendly schemes need to be designed to be cooperative with TCP flows by appropriately reacting to congestion [5]. There are a number of TCP-friendly congestion control algorithms which have recently been proposed, such as the rate-based Rate Adaptation Protocol (RAP) [18], equation-based TCP-Friendly Rate Control (TFRC) [6,7], and window-based Binomial Congestion Control (BCC) [19]. The transmission rates of the proposed TCP-friendly algorithms are generally smoother than that of TCP under stationary conditions at the expense of reduced responsiveness to changes in the network state (e.g., a new session arrival/departure to/from the bottleneck link) [20]. Moreover, these TCP-friendly mechanisms do not provide reliable transfer as TCP does, making them more suitable for real-time applications. The Datagram Congestion Control Protocol (DCCP) is a new transport protocol being developed by the IETF that provides a congestion-controlled flow of unreliable datagrams [21]. TCP-like congestion control without reliability and the equation-based TFRC [7] form the basis for the two congestion control profiles ID 2 and ID 3 in the DCCP protocol suite [22,23].

The stored video streaming problem over resource constrained networks, like the Internet, has attacted the attention of many researchers. Given network bandwidth and client buffer constraints, a dynamic programming algorithm with reportedly significant computational complexity is developed for the optimal selective frame discard problem in [10] as well as several heuristic algorithms. However, this study is unable to accomodate the bandwidth variability patterns of the Internet since the network bandwidth is assumed to be fixed and a priori known. On a similar ground, rate-distortion optimization-based video streaming algorithms have been developed in [24,25] that obtain scheduling policies for both new and retransmitted frames using stochastic control principles but the proposed methods are relatively complex and their feasability remain to be seen. The reference [26] considers a practical frame dropping algorithm for MPEG streams over best-effort networks but they neither use a TCP-friendly congestion control algorithm nor they take into account the deadlines of frames. In [27], a dynamic frame dropping filter for MPEG streams is proposed in a network environment where the available bandwidth changes dynamically but this work also lacks the TCP-friendliness component. A number of studies focus on streaming video using new TCP-friendly transport protocols [18,7] while others employing TCP itself [28–31]. One common objection to use of TCP for streaming applications is the fully reliable service model of TCP through retransmissions [30]. While delays due to retransmissions may not be tolerable for interactive applications, the service model for TCP may not be problematic for video on demand applications, which is the scope of

the current paper [30]. Moreover, the use of Explicit Congestion Notification (ECN) allows TCP to perform congestion avoidance without losses, limiting further the potential adverse effect of the TCP service model.

In this paper, we propose a stored video streaming system architecture which consists of an input buffer at the server side coupled with the congestion control scheme of TCP at the transport layer, for efficiently streaming stored video over the best-effort Internet. The proposed method can be made to work with other transport protocols including DCCP but our choice of TCP in the current paper as the underlying transport protocol stems from the following reasons:

- Slowly-responding TCP-friendly algorithms perform reasonably well in terms of video throughput in stationary conditions. However, responsiveness is especially critical in the core of the Internet today which appears to be operating in the transient rather than in the stationary regime due to the large session arrival and/or departure rates to/from the network. On the other hand, TCP congestion control has a well-established responsiveness to changing network state and might be more appropriate in rapidly changing environments.
- TCP with its original congestion control but with its full reliability feature replaced with selective reliability would be a more appropriate fit as a transport protocol for the underlying problem but the standards in this direction have not finalized and are still evolving [21,23]. We note that TCP's insistence on reliable delivery without timing considerations would adversely affect the performance of the system under packet losses especially for (near) real-time applications (e.g., applications requiring short playout delays). In this paper, we study the regimes for which TCP performance for stored video streaming is acceptable but also identify regimes for which TCP performs poorly and a new transport protocol would be needed.
- TCP is currently used for streaming applications in order to get through some firewalls that block UDP traffic.
- The choice of TCP as the transport protocol eliminates the unnecessary burden on the application-level designer by providing congestion control at the transport layer [21].
- Another key advantage related to providing congestion control at the transport layer (i.e., TCP) rather than "above UDP" is that the proposed scheme can make use of the services provided by the standard-based Explicit Congestion Notification (ECN) mechanism [32] which provides a means of explicitly sending a "congestion experienced" signal towards the TCP sender in TCP acknowledgment packets. We note that explicit feedback significantly reduces the losses in the network and is therefore particularly useful in scenarios such as video streaming where the frequency of retransmissions due to losses is to be kept at a minimum.

In our proposed architecture, the buffer management scheme selectively discards low priority frames from its head-end which otherwise would jeopardize the successful playout of high priority frames. Moreover, the proposed discarding policy is adaptive to changes in the bandwidth available to the video stream. Contrary to many of the previously proposed adaptive transmission algorithms, the proposed Selective Frame Discard (SFD) strategy is simple and easily implementable at the application layer by allowing additional information exchange between the transport layer and the application layer. Moreover, our proposed server-side frame discarding algorithm only needs to know the playout delay T_p and several network-related variables which are made available by using the services of TCP and the playout buffer occupancy does not need to fed back to the server in this proposed scheme. Our simulation results demonstrate that scalable stored video can efficiently be streamed over TCP with the proposed adaptive frame discarding strategy if the client playout delay is large enough to absorb the fluctuations in the TCP estimation of the available bandwidth. We also study the impact using Explicit Congestion Notification (ECN) in the network in terms of attained video quality. Finally, we compare the proposed edge-based server-side frame

discarding solution with the core-based Differentiated Services (Diffserv) Assured Forwarding (AF) Per-Hop-Behavior (PHB) architecture (see [33]) in the context of stored video streaming and identify regimes in which the former architecture outperforms the latter.

The rest of the paper is organized as follows. In Section 2, the proposed architecture including the scalable coding model and the selective frame discard schemes are presented. The simulation platform and the numerical results are given in Section 3. We conclude in the final section.

2. Video streaming architecture

In this section, we first describe our video encoding model and then present the details of the proposed input buffer management scheme based on selective frame discarding.

2.1. Scalable video coding

The main goal of scalable coding of video is to flexibly support a heteregoneous set of receivers with different access bandwidths and display capabilities. Furthermore, scalable coding provides a layered video bit stream which is amenable to prioritized transmission. In this paper, we assume that the stored video is encoded into two layers, the BL and the EL, using the Reference Picture Selection mode of H.263 version 2 [13,14]. In this structure (i.e., backward prediction disabled), the BL is composed of Intra (I) and anchor P (predicted) frames whereas the EL is composed of the remaining P frames. P frames in the EL are estimated using the anchor P frames or I frames in the BL where anchor P frames are chosen using the Reference Picture Selection mode. Throughout the rest of this paper, we will denote the base layer frames by H (High-priority), and enhancement layer frames as L (Low-priority). A schematic diagram of the employed scalable video coding structure is shown in Fig. 1. We leave the study of different temporal scalability models and other video coding standards for future research but we believe that the proposed architecture is applicable to other 2-layer scalable video codecs.

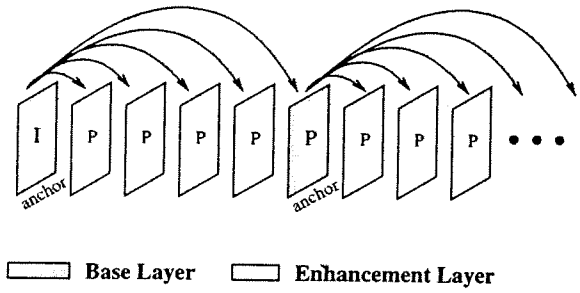


Fig. 1. Base and enhancement layers in temporal scalability mode.

2.2. Selective frame discarding

As stated in the previous section, we assume that video encoders generate H- and L-frames. If the available network bandwidth cannot accommodate the transmission of all frames, then it would be desirable to discard some of the L-frames on behalf of the H-frames. While making a L-frame discarding decision, our goal is to maximize the number of transported L-frames subject to the constraint that the loss rate for the H-frames would be minimal. In this definition, a loss refers to a missed frame at the client either because the frame is not transmitted by the server or is transmitted but partially/completely lost in the network or the frame is received by the client but after its deadline. For this purpose, we propose an input buffer implemented at the application layer of the sender which dynamically and intelligently discards L-frames from its headend and this scheme is depicted in Fig. 2.

We use the RTP/TCP/IP protocols stack in this study. We propose in this architecture that the stored video frames arrive at the input buffer at a frequency $f = 1/T$ frames per second, which is the frame generation rate of the underlying video session. These frames wait in the input buffer until they reach the headend of the buffer and a decision is then made by the Selective Frame Discard (SFD) block whether the corresponding frame should be passed towards the transport layer or is simply discarded. In cases of discard, the SFD block will make subsequent discard decisions until an acceptance decision is made. When a frame is accepted by the SFD module, it is segmented into

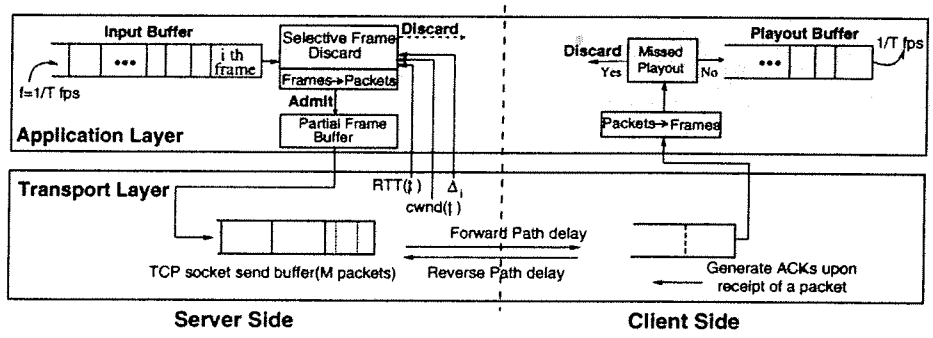


Fig. 2. Proposed stored video streaming architecture.

video packets (or RTP packets) of length at most L where we fix L to 1 Kbytes in this study. In our simulation studies, QCIF videos are encoded at around 30 dB quality and a typical video packet can carry 1–3 P-frames depending on the compression efficiency of the frame (i.e. high/low motion) and a typical I-frame can be transported by 2–3 video packets. Video packets of accepted frames are first placed in the partial frame buffer which is then drained by the TCP layer. We suggest that whenever a TCP packet begins to take its first journey towards the network, the TCP layer immediately retrieves a packet from the partial frame buffer if the buffer is nonempty. Otherwise, it queries the SFD module to make an acceptance/rejection decision on the head-end frame.

The acceptance/rejection decision is made as follows: The decision epoch for the ith frame is denoted by t_i irrespective of the outcome of the decision. The waiting time or the shaping delay in the input buffer for frame i, denoted by $D_{i,S}$, is the difference between t_i and the injection time for the ith frame to the input buffer. Let $D_{i,N}$ denote the network delay for the ith frame injected into the input buffer. Recalling that frames are generated by the encoder at integer multiples of T, the injection time for the ith frame to the input buffer will be $t_0 + iT$, where t_0 is the injection time of the 0th frame. The ith frame will then wait in the input buffer for $D_{i,S}$ seconds and the SFD module will make an admit/discard decision for the ith frame at time epoch $t_i \triangleq t_0 + iT + D_{i,S}$. If the ith frame is admitted by the SFD module into the transport layer then that frame will be delayed an additional $D_{i,\mathrm{TCP}}$ and $D_{i,N}$ seconds in the TCP buffer and in the network, respectively. It is clear that the ith frame must arrive at the receiver before its playout time $t_0 + D_{0,N} + T_p + iT$ where T_p is the initial buffering time of the playout buffer which starts accumulation as soon as the frame 0 arrives. So the following inequality should be satisfied for every accepted frame $i > 0$ for its succesful playout:

$$D_{i,S} \leqslant T_p - (D_{i,N} - D_{0,N}) - D_{i,\mathrm{TCP}} \quad (1)$$

In the above inequality, $D_{i,S}$ and T_p are known to the SFD module, however one needs to find estimates for the last two terms on the right hand side of the inequality. In this study, we suggest to estimate the one-way network delay difference $\Delta_i = D_{i,N} - D_{0,N}$ using the TCP Timestamps option (TSopt) in TCP headers [34]. In the TCP Timestamps Option, while transmitting packet m, the sender puts the transmission instant timestamp in the Timestamp Value (TSval) field. After receiving packet m, the receiver generates an acknowledgement packet denoted by ack m, by setting its *TSval* field with the current time of the receiver and by copying the *TSval* field of packet m to the Timestamp Echo Reply (TSecr) field of ack m. In this way, the SFD module will have an estimate of the one-way network delay difference using the TCP timestamp option for the last acknowledged TCP packet before time t_i when it needs to make a decision for frame i. On the other hand, the last term $D_{i,\mathrm{TCP}}$ is not known in advance but is relatively small compared to T_p unless there are TCP losses because of the mechanism described for initiating a data transfer from the application layer into the TCP layer. We therefore introduce a safety parameter α, $0 < \alpha < 1$ to account for the

errors due to inaccuracies due to estimations to be used in the inequality (1) as follows. In order for an admission decision for frame i to take place, the following new inequality should be checked by the SFD block:

$$D_{i,S} \leqslant \alpha(T_p - \Delta_i) \quad (2)$$

The inequality (2) can be used to select which frames to discard for nonscalable video but it needs to be modified for layered video. This modification is studied next.

2.3. Static and adaptive selective frame discard algorithms

We propose to use two different safety parameters α_L and α_H for the L-frames and the H-frames, respectively, for preferential treatment for H-frames. Such a treatment is possible by choosing $\alpha_L < \alpha_H$. This choice makes α_L not only a safety parameter but also a prioritization instrument. We summarize the general SFD algorithm at decision epoch t_i in Table 1.

The choice of the algorithm parameters α_L and α_H are key to the success of the proposed architecture. In Static SFD (SSFD), fixed α_L and α_H values are used throughout the video streaming session. However, such a fixed policy may not work well in all possible traffic scenarios. For example in cases where the instantenous available bandwidth is close to the the BL rate then the L-frames should aggressively be discarded (i.e., $\alpha_L \to 0$) in order to minimize the loss probability of the BL frames. On the other hand, if the available bandwidth happens to be close to or exceeds the total rate of the BL and the EL frames, then the L-frames should conservatively be discarded (i.e. $\alpha_L \to \alpha_H$). The very dynamic nature of the Internet may lead to significant variations in the available bandwidth even during the lifetime of a video session. Moreover

Table 1
The pseudo-code for the SFD algorithm at time t_i

```
if ((frame i == L-frame) && (D_{i,S} < α_L(T_p − Δ_i))){
  Admit();
} else if ((frame i == H-frame) && (D_{i,S} < α_H(T_p − Δ_i))){
  Admit();
} else Discard();
```

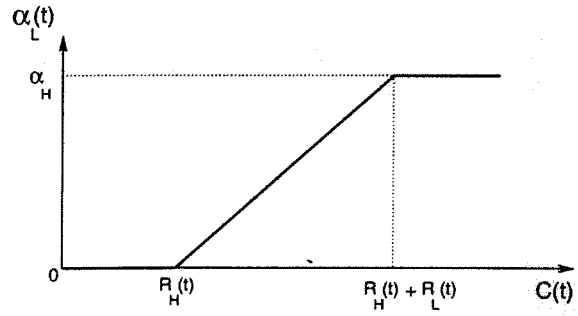


Fig. 3. Adaptive choice of α_L in the ASFD algorithm.

the instaneous BL and EL rates for VBR encoded video may substantially deviate from their long-run average values. These observations lead us to an adaptive version of the SFD algorithm. For this purpose, we define $C(t)$ as a smoothed estimate of the bandwidth available to the session at time t. Also we let $R_L(t)$ and $R_H(t)$ be the smoothed estimates of the EL and the BL, respectively, by monitoring the frame arrivals to the input buffer. We also let C, R_L and R_H denote the time averages of of the waveforms $C(t)$, $R_H(t)$, and $R_L(t)$, respectively. We then propose the simple Adaptive SFD (ASFD) scheme depicted in Fig. 3. We fix α_H and use it only as a safety parameter (α_H set to 0.7 in this study). The choice of α_L is less straightforward: α_L is zero when $C(t) < R_H(t)$, α_L equals α_H when $C(t) > R_H(t) + R_L(t)$ and it changes linearly within between these two end regimes. The notation SSFD(x) denotes the SSFD algorithm with $\alpha_H = 0.7$ and α_L set to x.

3. Simulation results

In this section, we study the performance of the proposed stored video streaming architecture using simulation. We use ns-2 [35] for simulations with a number of enhancements required for the video streaming architecture given in Fig. 2. We use the single bottleneck topology in Fig. 4 for all the simulation experiments. In all simulations, N video sessions (of length 780 s) share a single bottleneck link with capacity C_{tot} (set to 1 Mbps), where N will be varied to account for the variability of the available bandwidth to each user. The

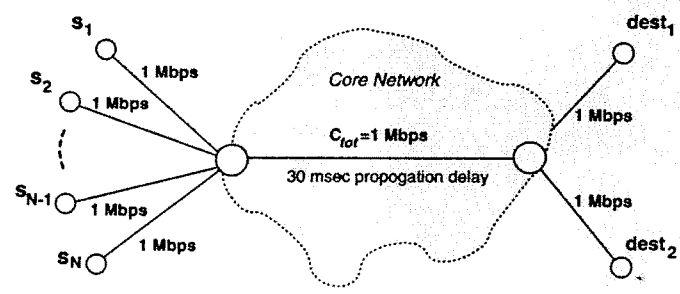


Fig. 4. The network topology used in the simulation studies.

buffer management mechanism for the bottleneck link is assumed to be Random Early Detect (RED). Motivated by [36], we use the RED parameters $(\min_{th}, \max_{th}, \max_p) = (20, 60, 0.1)$ and the RED smoothing parameter set to 0.002 unless otherwise stated.

The first $N/2$ sessions are sinked at $dest_1$ and the remaining ones at $dest_2$. Each video source employs TCP Reno with the same set of parameters and options and each source streams the same video clip. There is one tagged source we monitor among the N sources for Peak Signal-Noise Ratio (PSNR) plots. Each source starts streaming at random points in the video clip in order to prevent synchronization among the sources. Throughout the simulations, the bit rate of the VBR encoded video has substantial oscillations while the average rates are $R_L \approx 82.6$ kbps and $R_H \approx 35.0$ kbps (see Fig. 5). Given that the original video frequency is $f = 25$ frames/s, the two layer scalable video is composed of a single I and 9 anchor-P frames as the base layer for each two-seconds interval (i.e., Group of Pictures (GOP) duration). The remaining 40 are plain P frames that constitute the enhancement layer as given in Fig. 1. In our simulations, the average PSNR is used as the performance metric. Both the received frames and the lost frames are used in the PSNR calculation where the lost frames are concealed at the receiver by replicating the most recently decoded frame. Since we are using a temporally scalable bitstream, the PSNR of the received frames reflects the degradation in system performance due to losses only in the BL. By using PSNR for both received and lost frames as the performance metric, the degradation in the system performance caused by the L-frame losses are also included as well as the H-frame losses. In all of our experiments, the bottleneck

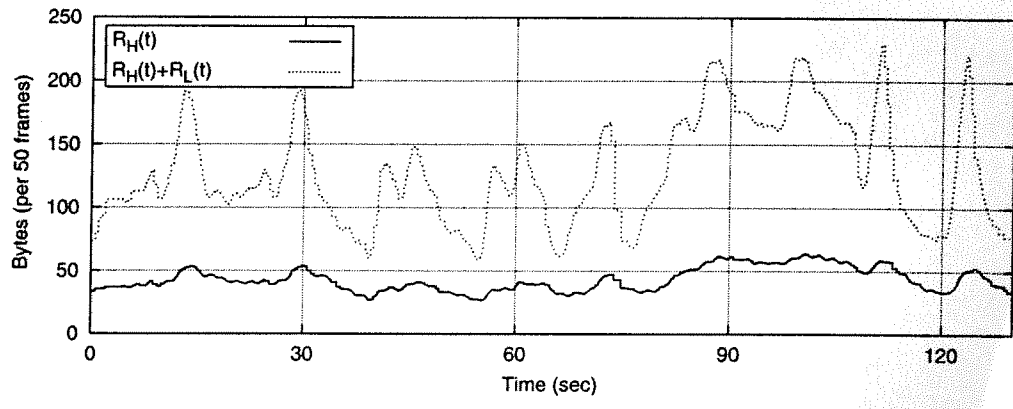


Fig. 5. Smoothed bit rates for the BL and EL for the layered video used in the simulations.

link with capacity C_{tot} is shared among N sources where $N \in \{6,\ldots,40\}$ and the expected fair bandwidth share per flow, which is $C \approx C_{tot}/N$, changes in the range $\{25,\ldots,166\}$ kbps.

In our first experiment, we compare and contrast the performance of the ASFD algorithm with the SSFD algorithm with three settings for $\alpha_L \in \{0.05, 0.4, 0.7\}$. For this purpose, we vary the number of video sessions N and thus change the fair share of each session $C \approx C_{tot}/N$ and obtain the corresponding PSNR value for the SSFD and ASFD algorithms. The playout delay T_p is set to 5 s in this study. The results are depicted in Fig. 6. The ideal curve is obtained by allowing the system to transmit and play all the scheduled frames, in other words for a given bandwidth it is assumed that there is enough playout buffering to tolerate the latency due to retransmissions and the video bitrate is properly matched to the constant available bandwidth in the network so that the scheduled frames never miss their playout times. In our simulations, the EL and/or BL frames are discarded sequentially for the computation of the ideal curve and the corresponding bitrate is calculated. The sequence used for discarding is the same for each GOP. The selection of a conservative SSFD policy (i.e., SSFD(0.05)) gives the best results for the heavy load case (i.e., $C < 100$ kbps) when compared to all other schemes. However, in the light load case when C gets close to or beyond $R_L + R_H$, the PSNR performance of SSFD(0.05) degrades substantially compared to the less conservative policies SSFD(0.4) and SSFD(0.7). On the other hand the adaptive version ASFD is robust with respect to the changes in the available bandwidth per user and it compares reasonably well with the best performing static policy in each case. The advantage of the ASFD is that the video server can find a policy very close to the optimal frame discarding policy using local measurements even when the available bandwidth per user changes significantly during the lifetime of the video session. This behavior can definitely not be obtained with static policies.

In our second simulation experiment, we study the impact of the RED parameters on the ASFD performance. The results are given in Fig. 7. The cases with three different RED configurations outperformed the drop-tail policy with the buffer size set to 120 packets. This observation can be explained by the fact that drop-tail buffer management causes synchronized losses and the resulting overshoots and undershoots in the resulting buffer occupancy yield substantial performance degradation relative to that of RED. We generally obtained quite robust results with RED but we also observed performance degradation with RED(10, 30, 0.1) in the heavy load case compared to the other two RED systems. This degradation is due to the relatively conservative choice of $\min_t$ and $\max_{th}$ in this system when a fairly large number of sources are multiplexed.

In the third simulation experiment, we study the impact of using ECN for which the RED modul

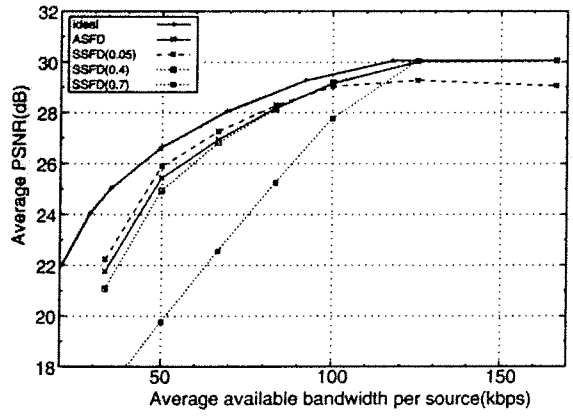


Fig. 6. Comparison of SSFD vs ASFD for the case $T_p = 5$ s.

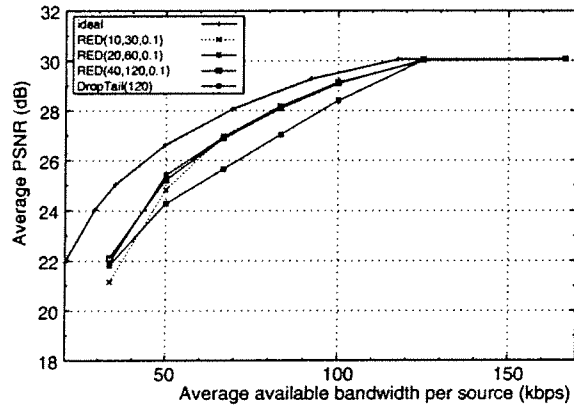


Fig. 7. Effect of RED parameters on ASFD performance with $T_p = 5$ s.

at the bottleneck link marks the packets with the corresponding probabilities as opposed to discarding them. This congestion information is then fed back in the TCP acknowledgements via which the TCP sources adjust their window sizes. Since all TCP senders are using ECN and all respond to congestion before actually loosing a packet, they tend to experience less the undesired data or timer driven loss recovery phases of TCP. This behaviour, as one might expect, leads to a significant performance improvement especially in congested network scenarios and for small initial playout delays. This situation is depicted in Fig. 8 in which T_p is set to 2 s and the performance of using TCP Reno without ECN and TCP Reno with ECN are shown in terms of the average PSNR values for varying C. For the heavy load case, the performance gain with ECN is remarkable (up to 2 db). The $T_p = 5$ s case is depicted in Fig. 9 for which the ECN gains are smaller compared to the $T_p = 2$ s case. For small playout delays, it is more likely that a larger percentage of the TCP's retransmissions arrive at the receiver later than their corresponding deadlines. With ECN, losses in the network are reduced and so are retransmissions. This is why the performance gain of ECN is more significant in cases with small playout delays. As shown in Fig. 8, $T_p = 2$ s of buffering cannot tolerate the timer driven retransmissions occuring in TCP, therefore a significant

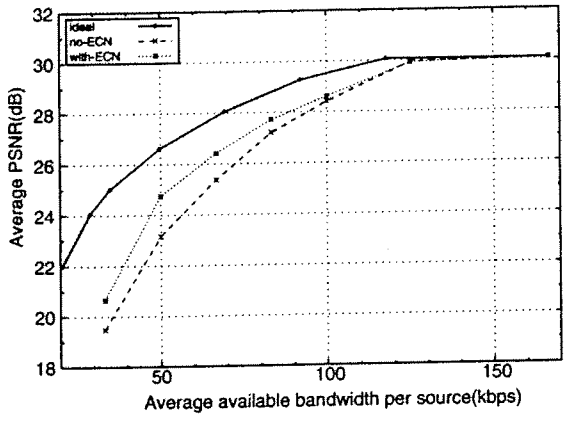


Fig. 8. Impact of ECN on streaming performance for ASFD with $T_p = 2$ s.

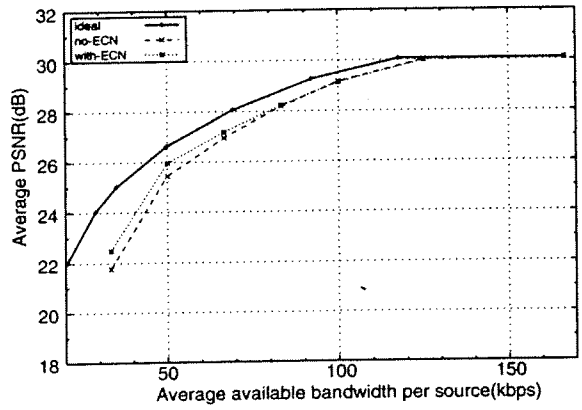


Fig. 9. Effect of ECN on streaming performance for ASFD with $T_p = 5$ s.

PSNR degredation is observed if ECN is not employed as compared to the $T_p = 5$ s case.

In the fourth experiment, we study the impact of the playout delay T_p which is used in order to compensate for the oscillations in the video bit rate and available network bandwidth per user. The playout delay T_p is varied from 1 s to 30 s and the corresponding PSNR values are plotted with respect to varying C in Fig. 10. The PSNR curves saturate at around $T_p = 15$ s beyond which buffering only slightly improves the PSNR performance. For small T_p (i.e., $T_p = 1$ s or 2 s), the playout delay is comparable to the delays encountered in TCP's data/timer driven retransmissions and a larger percentage of the network losses result in

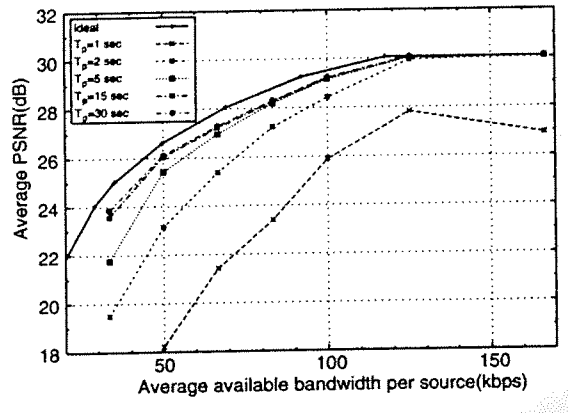


Fig. 10. Impact of T_p on average PSNR for ASFD algorithm.

missed playouts and thus reduced PSNRs. With TCP, increasing T_p from 2 to 5 s increases the streaming performance substantially by up to 3 dB.

Up to now, we assumed a best-effort Internet and we proposed intelligent frame scheduling and discarding techniques at the edge (i.e., at the application layer) which operates in harmony with the underlying transport protocol TCP. A network-based alternative for frame discrimination is the Internet Engineering Task Force (IETF) Differentiated Services (Diffserv) architecture [37]. Diffserv defines different service classes for applications with different Quality of Service (QoS) requirements. An end-to-end service differentiation is obtained by concatenation of per-domain services and Service Level Agreements (SLAs) between adjoining domains. Per domain services are realized by traffic conditioning including classification, metering, policing, shaping at the edge and simple differentiated forwarding mechanisms at the core of the network. One of the popular proposed forwarding mechanisms is Assured Forwarding (AF) Per Hop Behavior (PHB) [33]. The AF PHB defines four AF (Assured Forwarding) classes: AF1–4. Each class is assigned a specific amount of buffer space and bandwidth. Within each AF class, one can specify three drop precedence values: 1, 2, and 3. In the notation AFxy, x denotes the AF class number ($x = 1,\ldots,4$) and y denotes the drop precedence ($y = 1,\ldots,3$).

In our final simulation experiment, we compare the proposed edge-based server-side frame discarding solution with the core-based Differentiated Services (Diffserv) Assured Forwarding (AF) Per-Hop-Behavior (PHB) architecture in the context of stored video streaming and identify regimes in which the former architecture outperforms the latter. For the Diffserv scenario, we mark packets belonging to H-frames as AF11 and those of L-frames as AF12. We use Weighted RED (WRED) with the RED parameters (20,60,0.1) and (10,30,0.25) for AF11 and AF12, respectively [38]. We do not impose the use of any traffic conditioner in this experiment but we make use of only the differentiated forwarding paradigm of Diffserv. We use User Datagram Protocol (UDP) for the transport layer for this scenario. We will refer to the combined scheme as Diffserv + UDP. The number of video sources sharing the bottlenk link are varied and PSNR values are plotted in Fig. 11 for the case $T_p = 1$ s which demonstrates that when the client playout delay T_p is small and comparable to one Round Trip Time (RTT), the Diffserv+UDP solution outperforms the proposed ASFD+TCP approach. However, when T_p is increased to 5 s, then the ASFD+TCP solution gives better results than that of the Diffserv+UDP solution (see Fig. 12). The reason for this behaviour is that when the client playout delay is large enough then the TCP sender can retransmit not ACKed packets without them missing their deadlines (as opposed to the $T_p = 1$ s case). Moreover, it is the application layer that intelligently decides

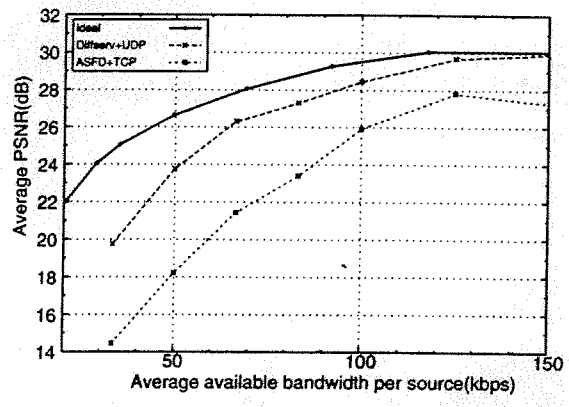


Fig. 11. PSNR plots using Diffserv + UDP and ASFD + TCP scheme for $T_p = 1$ s scenario.

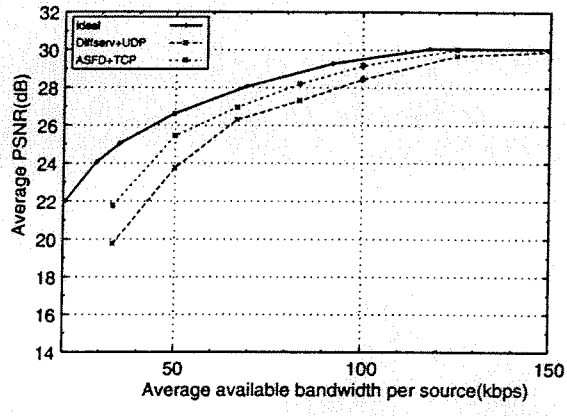


Fig. 12. PSNR plots using Diffserv+UDP and ASFD+TCP scheme for $T_p = 5$ s scenario.

on which frames to discard in ASFD + TCP by taking into consideration their playout deadlines. We're led to believe that when the playout delays are sufficiently large (i.e., $T_p > 5$ s) then the proposed edge-based adaptive approach is superior to the network-based Diffserv+UDP scheme which is static in its parameter settings and which is not aware of the playout deadlines.

4. Conclusions

Motivated by the extensive operation experience behind TCP, we propose in this paper an easily implementable stored video streaming system using TCP transport. The proposed system consists of an input buffer implemented at the application layer of the server coupled with the congestion control scheme of TCP at the transport layer. The proposed frame discarding strategy dynamically and intelligently discards low priority frames from its head-end. Moreover, it is adaptive to changes in the bandwidth available to the video stream. Our simulation results demonstrate that scalable stored video can efficiently be streamed over TCP with the proposed adaptive frame discarding strategy if the client playout delay is large enough to absorb the fluctuations in the TCP estimation of the available bandwidth. As expected, the use of Explicit Congestion Notification (ECN) in the network is shown to slightly improve the throughput especially in congested network scenarios and for small initial playout delays. Finally, we compare the proposed edge-based server-side frame discarding solution with the core-based Differentiated Services (Diffserv) AF PHB architecture and identify regimes in which the former architecture outperforms the latter. We show through a number of simulations that if the playout delay is sufficiently long (i.e., $T_p > 5$ s) then the proposed edge-based solution outperforms the core-based Diffserv solution whereas this relationship is reversed otherwise.

References

[1] D. Wu, Y.T. Hou, Y.Q. Zhang, Transporting real-time video over the Internet: Challenges and approaches, Proc. IEEE 88 (12) (2000) 1855–1875.

[2] M. Civanlar, A. Luthra, S. Wenger, W. Zhu, Introduction to the special issue on streaming video, IEEE Trans. Circuits Syst. Video Technol. 11 (3) (2001) 265–268.

[3] N. Laoutaris, I. Stavrakakis, Intrastream synchronization for continuous media streams: A survey of playout schedulers, IEEE Network 16 (3) (2002) 30–40.

[4] M. Kalman, E. Steinbach, B. Girod, Rate-distortion optimized video streaming with adaptive playout, in: Proceedings of ICIP, Vol. 3, Rochester, NY, 2002, pp. 189–192.

[5] S. Floyd, K. Fall, Promoting the use of end-to-end congestion control in the Internet, IEEE/ACM Trans. Networking 7 (4) (1999) 458–472.

[6] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, Modeling TCP Reno performance: A simple model and its empirical validation, IEEE/ACM Trans. Networking 8 (2) (2000) 133–145.

[7] S. Floyd, M. Handley, J. Padhye, J. Widmer, Equation-based congestion control for unicast applications, in: ACM SIGCOMM, Stockholm, Sweden, 2000, pp. 43–56.

[8] A. Lippman, Video coding for multiple target audiences, in: SPIE Conference on Visual Communications and Image Processing, Vol. 3653, 1999, pp. 780–782.

[9] G.J. Conklin, G.S. Greenbaum, K.O. Lillevold, A.F. Lippman, Y.A. Reznik, Video coding for streaming media delivery on the Internet, IEEE Trans. Circuits Syst. Video Technol. 11 (3) (2001) 269–281.

[10] Z.-L. Zhang, S. Nelakuditi, R. Aggarwal, R.P. Tsang, Efficient selective frame discard algorithms for stored video delivery across resource constrained networks, in: INFOCOM, Vol. 2, 1999, pp. 472–479.

[11] B.G. Haskell, A. Puri, A.N. Netravali, Digital Video: An Introduction to MPEG-2, Kluwer Academic Publishers, Boston, MA, 1996.

[12] A. Puri, T. Chen, Multimedia Systems, Standards, and Networks, Marcel Dekker, New York, 2000.

[13] Video coding for low bit rate communication, ITU-T Recommendation H.263 (February 1998).

[14] G. Cote, B. Erol, M. Gallant, F. Kossentini, H.263+: video coding at low bit rates, IEEE Trans. Circuits Syst. Video Technol. 8 (7) (1998) 849–866.

[15] A. Luthra, G.J. Sullivan, T. Wiegand, Introduction to the special issue on the H.264/AVC video coding standard, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 557–559.

[16] M. Ghanbari, Layered coding, in: M.T. Sun, A.R. Reibman (Eds.), Compressed Video Over Networks, Marcel Dekker, New York, 2001, pp. 251–308.

[17] H. Radha, M. vanderSchaar, Y. Chen, The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP, IEEE Trans. Multimedia 3 (1) (2001) 53–68.

[18] R. Rejaie, M. Handley, D. Estrin, RAP: An end-to-end rate-based congestion control mechanism for realtime streams in the Internet, in: Proceedings of INFOCOM, Vol. 3, 1999, pp. 1337–1345.

[19] D.Bansal, H. Balakrishnan, Binomial congestion control algorithms, in: Proceedings of INFOCOM, Vol. 2, 2001, pp. 631–640.
[20] Y. Yang, M. Kim, S. Lam, Transient behaviors of TCP-friendly congestion control protocols, in: Proceedings of INFOCOM, 2001, pp. 1716–1725.
[21] E. Kohler, M. Handley, S. Floyd, Datagram congestion control protocol (DCCP), Internet draft draft-ietf-dccp-spec-09.txt, work in progress, November 2004.
[22] S. Floyd, E. Kohler, J. Padhye, Profile for DCCP congestion control ID3: TFRC congestion control, IETF Internet-draft draft-ietf-dccp-ccid3-09.txt, November 2004.
[23] S. Floyd, E. Kohler, Profile for DCCP congestion control ID2: TCP-like congestion control, IETF Internet-draft draft-ietf-dccp-ccid2-08.txt, November 2004.
[24] M. Podolsky, S. McCanne, M. Vetterli, Soft ARQ for layered streaming media, Tech. Rep. UCB/CSD-98-1024, University of California, Computer Science Division, Berkeley, November 1998.
[25] P. Chou, Z. Miao, Rate-distortion optimized streaming of packetized media, Tech. Rep. MSR-TR-2001-35, Microsoft Research, February 2001.
[26] M. Hemy, U. Hengartner, P. Steenkiste, MPEG systems in best-effort networks, in: Packet Video Workshop, New York, 1999.
[27] H. Cha, J. Oh, R. Ha, Dynamic frame dropping for bandwidth control in MPEG streaming system, Multimedia Tools Appl. 19 (2003) 155–178.
[28] Y. Dong, R. Rakshe, Z.-L. Zhang, A practical technique to support controlled quality assurance in video streaming across the Internet, in: International Packet Video Workshop, Pittsburgh, Pennsylvania, USA, 2002.
[29] P. Mehra, A. Zakhor, TCP-based video streaming using receiver-driven bandwidth sharing, in: International Packet Video Workshop, Nantes, France, 2003.
[30] C. Krasic, K. Li, J. Walpole, The case for streaming multimedia with TCP, in: 8th International Workshop on Interactive Distributed Multimedia Systems (iDMS 2001), 2001.
[31] I.V. Bajic, O. Tickoo, A. Balan, S. Kalyanaraman, J. Woods, Integrated end-end buffer management and congestion control for scalable video communications, in: International Conference on Image Processing, Vol. 3, 2003, pp. 257–260.
[32] S. Floyd, TCP and explicit congestion notification, ACM Comput. Commun. Rev. 24 (5) (1994) 10–23.
[33] J.Heinanen, F. Baker, W. Weiss, J. Wroclawski, Assured forwarding PHB group, RFC 2597, IETF, June 1999.
[34] S. Floyd, TCP extensions for high performance, RFC 1323, IETF (May 1992).
[35] UCB/LBNL/VINT, The Network Simulator - ns-2. URL `http://www.isi.edu/nsnam/ns/`.
[36] S. Floyd, V. Jacobson, Random early detection gateways for congestion avoidance, IEEE/ACM Trans. Networking 1 (4) (1993) 397–413.
[37] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An architecture for differentiated service, RFC 2475, IETF, December 1998.
[38] U. Bodin, U. Schelen, S. Pink, Load-tolerant differentiation with active queue management, ACM Computer Communication Review 30 (3) (2000) 4–16.

Eren Gürses received the B.S. and M.S. degrees from Middle East Technical University, Turkey, in 1996 and 1999, respectively, in Electrical and Electronics Engineering where he is currently pursuing his Ph.D. degree. His current research interests are multimedia communications over packet networks and rate adaptive video coding.

Gozde Bozdagi Akar received the B.S. degree from Middle East Technical University, Turkey, in 1988 and M.S. and Ph.D. degrees from Bilkent University, Turkey, in 1990 and 1994, respectively, all in electrical and electronics engineering. She was with the University of Rochester and Center of Electronic Imaging Systems as a visiting research associate from 1994 to 1996. From 1996 to 1998, she worked as a member of research and technical staff at Xerox Corporation—Digital Imaging Technology Center, Rochester. From 1998 to 1999 she was with Baskent University, Department of Electrical and Electronics Engineering. During the summer of 1999, she worked as a visiting researcher at the Multimedia Labs of NJIT. Currently, she is an Associate Professor with the Department of Electrical and Electronics Engineering, Middle East Technical University. Her research interests are in video processing, compression, motion modeling and multimedia networking.

Nail Akar received the B.S. degree from Middle East Technical University, Turkey, in 1987 and M.S. and Ph.D. degrees from Bilkent University, Turkey, in 1989 and 1994, respectively, all in electrical and electronics engineering. From 1994 to 1996, he was a visiting scholar and a visiting assistant professor in the Computer Science Telecommunications program at the University of Missouri-Kansas City. In 1996, he joined the Technology Planning and Integration group at the Long Distance Division, Sprint, Kansas, USA, where he held a senior member of technical staff position from 1999 to 2000. Since 2000, he is an assistant professor at Bilkent University. His current research interests include performance analysis of computer and communication networks, queueing systems, traffic engineering, network control and resource allocation, and multimedia networking.

Packet Video 2003

SELECTIVE FRAME DISCARDING FOR VIDEO STREAMING IN TCP/IP NETWORKS

Eren Gürses, Gözde Bozdağı Akar

Dept. of Electrical and Electronics Eng.
Middle East Technical Univ, Ankara, Turkey
gurses,bozdagi@eee.metu.edu.tr

Nail Akar

Dept. of Electrical and Electronics Eng.
Bilkent University, Ankara, Turkey
akar@ee.bilkent.edu.tr

ABSTRACT

TCP (Transmission Control Protocol) with its well-established congestion control algorithm is the prevailing transport layer protocol for non-real time data in current IP (Internet Protocol) networks. It would be desirable to transmit any type of multimedia data using a variant of TCP in order to take advantage of the extensive operational experience behind TCP in the Internet. However, some features of TCP including retransmissions and variations in throughput and delay, although not catastrophic for non-real time data, may result in inefficient video transport if not properly engineered. There are a number of proposals that modify TCP in order to efficiently transport stored video. In this paper, we propose an architecture which includes an input buffer at the server coupled with the congestion control scheme of TCP at the transport layer. This buffer selectively discards low priority frames from its head-end which otherwise would jeopardize the successful playout of high-priority frames. This architecture is applied to several TCP variants and our results demonstrate that scalable stored video can efficiently be transmitted over IP networks if the client buffering time is long enough to absorb the fluctuations in the estimated network bandwidth arising due to TCP. Moreover, certain retransmission strategy modifications to TCP/Reno are shown to significantly improve the performance.

1. INTRODUCTION

The transmission of high quality video over the Internet is now becoming a reality due to recent progresses in video compression and networking technologies, efficient video coders/decoders and increasing interest in applications such as video on demand, videophone, and video conferencing. TCP (Transmission Control Protocol) with its well-established congestion control algorithm is the prevailing transport layer protocol for non-real time data in the current Internet. However, some features of TCP, such as retransmissions and variations in throughput and delay, are generally believed to be unsuitable for transporting video. For streaming video, User Datagram Protocol (UDP) is often used. However UDP does not have a built-in congestion control mechanism, so most of the video streams are unable to respond to network congestion and this adversely affects the network performance as a whole. Potential for future congestion collapse of the Internet due to flows that do not use responsible end-to-end congestion control is addressed in [1]. There are a number of options to address this problem in the context of stored video streaming [2]: (a) employing congestion control above UDP, (b) using congestion control at the transport layer using a new protocol designed from scratch for unreliable data flows, (c) using congestion control at the transport layer using a suitable modification of a well established standard such as TCP. A key disadvantage of providing congestion control above UDP is that it places an unnecessary burden on the application-level designer. A second issue related to providing congestion control above UDP is that it would require giving up the use of Explicit Congestion Notification (ECN) [3] which would otherwise provide a mechanism to explicitly feedback a "congestion experienced" signal on transport layer acknowledgment packets. We note that explicit feedback minimizes the losses in the network and is therefore particularly useful in scenarios like video streaming where the frequency of retransmissions due to losses is to be kept at a minimum. In this paper, we pursue the path of providing congestion control at the transport layer. However, as opposed to designing a new transport protocol for unreliable data flows (which is outside the scope of this paper but we refer the reader to [2] for a detailed discussion), we study the performance of a selective frame discard strategy implemented at the server side which is tightly coupled to a transport layer implemented using existing modifications to the well-known TCP suite. These modifications do not nullify retransmissions but instead minimize the retransmission timeouts and their adverse affect on video quality. This architecture is applied to several TCP variants. Our results demonstrate that scalable stored video can efficiently be transmitted over IP networks if the playout time is long enough to absorb the fluctuations in estimated network bandwidth arising due to TCP.

The rest of the paper is organized as follows: In Section 2, we present a summary of related work. In Section 3, the proposed architecture including the scalability and selective frame discard schemes are given. Section 4 covers the background for TCP variants addressing their drawbacks for video delivery together with proposed modifica-

tions. The simulation platform and results are presented in section 5. We conclude in the final section.

2. RELATED WORK

Recently, a number of methods have been proposed for congestion control and quality adaptation for Internet video streaming [4, 5, 6, 7]. In most of these studies, congestion control is performed in the application layer. In [6], Feamster et. al. discuss several quality adaptation schemes using binomial congestion control for video streaming using RTP/UDP. The feedback to be used by the congestion manager [8] is obtained using RTCP. The results show that quality improvements can be obtained if hierarchically encoded video is transmitted using an input buffer management system. In [9], Sisalem and Schulzrinne present a rate adaptation algorithm for multimedia applications. Their work also depends on RTP/RTCP. Feng et. al. propose an adaptive smoothing mechanism in [10] by dropping low priority frames.

In [5], Hsiao et al. address the congestion control problem for Internet video streaming using a modification of TCP. Their algorithm avoids congestion by delaying acknowledgment (ACK) packet generation at the receiver based on the congestion notification from the routers. Even though the results are promising, real world implementation is a problem due to the required modification of the receiver protocol stack. Rajaie et. al. present a quality adaptation mechanism in [4] for video streaming while using a TCP-friendly congestion control mechanism. The presented algorithm includes a linear allocation of available bandwidth among different layers of video. In [7], Balan et. al. present an integrated scheme based on interworking between live adaptive encoding, packet filtering at the sender and TCP-friendly binomial congestion control scheme. The packets are dropped by using the priority information from the encoder and the network information from the congestion control scheme. In [11] Saparilla and Ross found optimal rates that should be allocated to each layer in a two layer video, using receiver feedbacks about the playout buffer information.

3. PROPOSED ARCHITECTURE

In this section, we first describe our video encoding model and then give the details of the proposed input buffer model based on selective frame discarding.

3.1. Scalable Video Coding

The main goal of scalable coding of video is to flexibly support multiple receivers with different access bandwidths, display capabilities and display requests to allow video database browsing and multiresolution playback of video content in multimedia environments. Another goal of scalable coding is to provide a layered video bit stream which is amenable for prioritized transmission. Many scalable video-coding techniques have been proposed over the past few years for real-time Internet applications by several video compression standards such as MPEG-2/4 and H.263/263+ [12]. The types of scalability which are defined in these standards can be categorized as temporal, spatial, SNR, object (only for MPEG4) scalability. All these types of scalable video consist of a Base Layer (BL) which is the minimum amount of data needed for decoding the video stream and one or more Enhancement Layers (EL). The EL part of the stream represents additional information. Both the base layer and the enhancement layer can be composed of I-P-B (Intra-Inter (Predicted-Bidirectionally predicted)) pictures which are the three generic picture types used in the above-mentioned standards. A schematic diagram of scalable video coding using temporal scalability is shown in Figure 1. In this figure, the base layer is composed of the I and P pictures whereas the enhancement layer is composed of P pictures. Other than the temporal scalability, SNR scalability is also widely used. One of the drawbacks of this approach is that when one of the EP frames (Enhancement Layer-P frame) is lost, the EP's quality will drastically degrade. Another scalability structure is *Fine Granular Scalability* (FGS). In FGS, there is no temporal relation among the frames in the EL. Since in FGS the EL is formed of bit-plane blocks which are DCT coded, bandwidth may be utilized more efficiently. However because of lack of temporal relation, increase in bit rate occurs especially in cases where the BL bit rate is chosen to be small as compared to the total rate. In order to solve the above-mentioned problems, we used the Reference Picture Selection mode of H.263+ (Annex N) [13] in this work. This is a simpler version of the temporal scalability mode of H.263+ (Annex O), with backward prediction disabled.

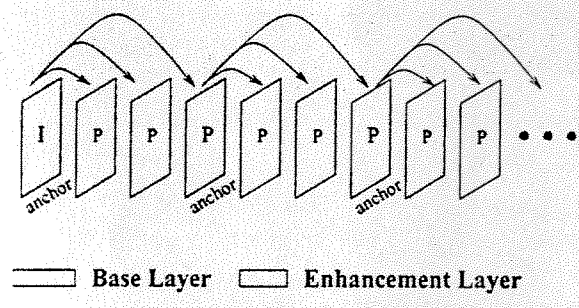


Fig. 1. Base and enhancement layers.

Throughout the rest of this paper we will denote the base layer frames as H (high priority), and enhancement layer frames as L (low-priority).

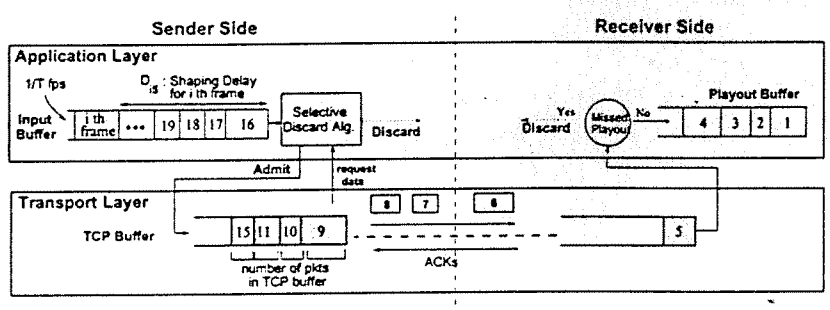


Fig. 2. Architecture of the proposed system

3.2. Selective Frame Discarding

As stated in the previous section, video encoders produce frames of different importance. If the available network bandwidth cannot accommodate all frames then it is desirable to discard the low priority frames when needed in behalf of the high priority ones. Such a selective treatment can best be carried out at the congestion location in the network but this requires differentiated services [14] support of the network(s) between the sender and the receiver. In the foreseeable future, we do not expect full-scale deployment of differentiated services and therefore we suggest preferential treatment for high-priority frames at the application layer of the sender. We assume that the available bandwidth is sufficient for transporting the high priority frames and our goal is to maximize the number of transported low priority frames subject to the constraint that the loss rate for the high priority frames would be minimal (if not zero). For this purpose, we propose an input buffer implemented at the application layer of the sender which, from its head-end, dynamically discards low priority frames based on an algorithm described below.

Figure 2 depicts the architecture of the proposed system. Stored video frames of both base and enhancement layer arrive at the input buffer at integer multiples of T, dictated by the actual video bit rate. When the transport layer requests new data from the input buffer, a decision is to be made whether to hand the foremost frame over to the transport layer or discard it. Since the input buffer is drained at a variable rate governed by the congestion control algorithm of the transport layer, the interdeparture times from the input buffer between two consecutive frames will be said to be *shaped*. In Figure 2, D_{iS} and D_{iN} denote the shaping delay and the network delay, respectively, of the $i^{th}, i = 0, 1, 2, \ldots$ frame injected into the input buffer. Considering that frames are generated by the encoder at integer multiples of T, the generation time for the i^{th} frame will be $t_0 + iT$, where t_0 is the generation time of the frame zero. The i^{th} frame will then wait in the input buffer for a time D_{iS} and if admitted into the transport layer, will be delayed in the network for another D_{iN} sec. At this point it should be noted that, admitted frames will be wrapped with proper TCP headers and in the rest of the transport layer operations they will be called as *packets*. At this step in order to prevent IP fragmentation in the network layer, larger frames are fragmented into packets of at most 1000 bytes at the transport layer and then wrapped with appropriate TCP headers. In order to allow variable size TCP packets, which are fixed sized segments in the original TCP implementations, accepted amount of data to the network during a round-trip time (i.e. congestion window) for each TCP source is calculated in terms of bytes rather than segments.

Table 1. Frame generation (sender) and playout (receiver) instances. Input buffer shaping delay + propagation & queuing delay in the network

Frame	Generation Time (sender)	Shaping+N/W Delay	Playout Time (receiver)
0	t_0	$D_{0S} + D_{0N}$	$t_0 + D_{0N} + T_p$
1	$t_0 + T$	$D_{1S} + D_{1N}$	$t_0 + T + D_{0N} + T_p$
		.....	
i	$t_0 + iT$	$D_{iS} + D_{iN}$	$t_0 + iT + D_{0N} + T_p$

At the receiver's transport layer these packets are received and directly passed to the application layer in order to reassemble and generate the frames. Frames must be played out at the receiver at times $t_0 + iT + D_{0N} + T_p$ where T_p is the initial buffering time of the playout buffer, which starts after the frame zero completely arrives (Table 1). Since the frame i arrives at the playout buffer at $t_0 + iT + D_{iS} + D_{iN}$, we have to satisy the following inequality to achieve a smooth playout:

$$t_0 + iT + D_{iS} + D_{iN} \leq t_0 + iT + D_{0N} + T_p \quad (1)$$

$|D_{0N} - D_{iN}|$ is bounded by the maximum delay jitter in the network denoted by $J_{max} = (1 - \alpha)T_p$. We assume that the parameter α lies in the interval $(0, 1)$ since in properly dimensioned applications, T_p should be chosen greater than the maximum network jitter. Therefore, we reduce the inequality (1) to the following:

$$D_{iS} \leq \alpha T_p, \; 0 \leq \alpha \leq 1 \quad (2)$$

Algorithm 1 gives the pseudo-code for the drop policy where

Table 2. ALGORITHM 1 - Algorithm for source buffer management where $frame_i$ is at the head of the buffer

```
if {priority(frame_i) == L) {
   if ( D_iS ≤ α T_p)
      admit(frame_i);
   else
      discard(frame_i);
} else {
   admit(frame_i);
}
```

the frames are dropped based on inequality (2) . Since high priority frames which form the base layer should be delivered to the receiver without any loss, this drop policy is only applied to low priority (enhancement layer) frames. Two important factors in this algorithm is the choice of α and T_p. Detailed analysis of the impact of these parameters on the video streaming performance is presented in Section 5.

Algorithm 1 gives the pseudo-code for the drop policy where the frames are dropped based on inequality (2) . Since high priority frames which form the base layer should be delivered to the receiver without any loss, this drop policy is only applied to low priority (enhancement layer) frames. Two important factors in this algorithm is the choice of α and T_p. Detailed analysis of the impact of these parameters on the video streaming performance is presented in Section 5.

4. IMPROVEMENTS ON LOSS RECOVERY/DETECTION

TCP with its well-established congestion control algorithm is the prevailing transport layer protocol for non-real time data in the current Internet. However, some features of TCP including retransmissions and variations in throughput and delay, although not catastrophic for non-real time data, may result in inefficient video transport if not properly engineered. In the remainder of this section, we will describe the existing TCP options and flavors, namely Limited Transmit [15], Retransmitted Packet Loss Detection (RPLD) [19], Selective Acknowledgments (SACK) [16], Forward Acknowledgments (FACK) [17], Explicit Congestion Notification (ECN) [3] for efficient video streaming. With these options, a TCP sender achieves two main goals: *(i) Improving Loss Recovery/Minimizing Loss, (ii) Improving Loss Detection.* Also it should be noted that these two goals should not be mutually exclusive, since without an appropriate loss recovery, improving detection of packet loss becomes useless, and vice versa.

The proposed methods above, are used to make the TCP utilize the data-driven recovery in case of a packet loss,

Table 3. Methods of Reducing RTO

Improve Loss Recovery	Minimize Loss in N/W
- Using SACK - Using RPLD	- Using ECN

rather than timer-driven methods. However these methods improve only the inefficiency of TCP caused by retransmissions. Other undesired features of TCP, such as variation in rate and delay, may be reduced by using congestion control algorithms with smoother rate as given in Section 4.3 at the expense of slow responsiveness.

4.1. Improving Loss Recovery / Minimizing Loss in N/W

The loss recovery phase of a TCP sender has 2 tasks; (i) attempting to recover from losses (ii) responding to the congestion in the network by decreasing the *cwnd* (congestion window). There are basically two types of loss recovery mechanisms in TCP, namely *data-driven* and *timer-driven* mechanisms. *Fast Transmit* and *Fast Recovery* are data-driven [18], and *Retransmission Timeout (RTO)* is a timer-driven loss recovery method of TCP [15]. Recovering losses with an RTO is not a desired way of loss-recovery in video streaming applications since it causes the TCP sender to stay idle for an extended period of time which is undesirable for video applications. Methods of reducing the frequency of RTOs is given in Table 3. The first approach is improving the loss recovery characteristic of a standard TCP sender by means of introducing some extra signaling (i.e., SACK) or using extra information as in RPLD. In the second approach, losses causing RTOs in the network can be reduced by congested routers marking a "congestion experienced" field using the standard-based ECN field [3]. Receivers then convey this information back to the sender using the transport layer acknowledgment packets upon the receipt of which, senders can adjust their windows before too late. When there are fewer losses in the network, fewer loss recoveries will occur, resulting in reduced frequency of RTOs.

4.1.1. TCP-SACK

TCP receivers implementing SACK [16] use an extra field for the three most recently received acknowledged segment-block information, in addition to the sequence number of the last ACKed packet (without any hole) which is common to all TCP variants (Tahoe, Reno, New Reno, SACK, FACK). By this way, the sender is informed about the actually received segments by the receiver. This enables the protocol not to retransmit the already received packets in either the timer or the data-driven loss recovery methods. We note that TCP/SACK receivers are quite popular and implemented in the TCP/IP stacks of all new operating systems.

4.1.2. RPLD

If a retransmitted TCP packet is lost, the sender does not make any data-driven loss recovery attempt to recover from the loss and RTO eventually occurs. Therefore, implementation of a mechanism which handles losses in the retransmitted packets will be beneficial for video streaming type of applications. In the implementation of RPLD [19], a loss detection algorithm (which is implemented on the sender side) puts the local transmission time information for each transmitted segment in its *transmitted packets history list*. When an ACK (with SACK fields) is received by the sender, it checks all of the sacked segments together with the local transmission time information (from the transmitted packets history list) and detects retransmitted packet losses. On the detection of retransmitted packet losses, they are immediately retransmitted again if *cwnd* allows.

4.1.3. Explicit Congestion Notification (ECN)

Traditionally a router reacts to congestion by dropping a packet in the absence of buffer space, which is called a "tail drop". Such a drop is implicitly disseminated towards the sender via acknowledgments. The sender then adapts its rate based on the underlying TCP congestion control implementation. In such a configuration, packet loss is one of the basic building blocks of congestion control and is inevitable. Recently, improved congestion control mechanisms in routers have emerged. One such mechanism is Random Early Detection (RED) which detects incipient congestion and implicitly signals the oversubscribing flow to slow down by dropping its packets. A RED-enabled router detects congestion before the buffer overflows, based on a running average queue size, and drops packets probabilistically before the queue actually fills up. An extension to RED is to mark a certain "congestion experienced" field in the IP header rather than dropping packets. Cooperating end systems would then use this as a signal that the network is congested and slow down. This is known as Explicit Congestion Notification (ECN) and ECN can be very effective in reducing losses in the network. However, we note that active network involvement is required for ECN and deployment of ECN-capable systems are in their early stages of evolution.

4.2. Improving Loss Detection

In order to make a loss recovery, loss should be first detected using TCP's negative acknowledgments (ACKs). TCP is an ACK-clocked protocol, therefore the TCP receiver should generate enough ACKs for the TCP sender in order to detect a loss (i.e. congestion) and respond without waiting for an RTO. Therefore, improved and more robust loss detection is directly related to the continuity of ACK clocking. Improved calculation of outstanding packets in the network, and ability to work even with small and large congestion windows, will provide the continuity of ACK clocking.

Making ack-clocking mechanism better doesn't remove RTOs however it just enables the protocol to detect losses better. But if loss recovery is not improved (i.e. using RPLD) or number of losses is minimized (i.e. using ECN), RTOs will still occur.

4.2.1. FACK (Forward Acknowledgment)

TCP/FACK [17] uses the additional SACK information to keep an explicit measure of the total number of TCP segments outstanding in the network. Therefore TCP/FACK sender is implemented in conjunction with a TCP/SACK receiver. TCP/FACK improves the ACK clocking performance because it estimates the number of packets in the network more accurately than TCP/Reno and TCP/SACK by using the state variables $snd.nxt$ and $retran\_data$. snd_next is incremented after transmitting a new segment. On the other hand $retran\_data$ is incremented after retransmitting a lost segment in a data-driven loss recovery (Fast Retransmit & Fast Recovery) Phase. The maximum Sacked segment number is held in the $snd.fack$ variable, and these 3 variables together with the *cwnd* (congestion window) are used to determine whether or not to send a segment (new segment or a retransmitted one).

$$while(snd.nxt < snd.fack + cwnd - retran\_data)$$
$$\quad sendsomething()$$

Better estimation of outstanding packets in the network results in an improved ack-clocking mechanism. Additionally, extracting the information of three duplicate acknowledgments (dupacks) from the SACK fields of the ACKs in order to detect three dupacks, rather than waiting for the arrival of three separate ACK packets with duplicate sequence numbers, also improves the ack-clocking property.

Using a TCP/FACK sender requires a standard TCP/SAC receiver. Therefore TCP/FACK implementation requires a change only in the TCP/IP stack of the sender, not the receiver.

4.2.2. Limited Transmit

Since TCP is an ACK-clocked (or self-clocked) mechanism, the TCP receiver should generate enough ACKs for the TCP sender in order to detect losses due to congestion and respond accordingly without having to wait for an RTO. "Limited Transmit" [15] provides the continuity of self-clocking by increasing the *cwnd* artificially for each incoming dupack, in cases when the $cwnd \leq 3$. In situations when the

number of outstanding packets in the network is less than or equal to three, and if a packet is dropped the receiver would never receive three dupacks and cannot make a data-driven "loss recovery".

4.3. Smoothing the Rate

Binomial congestion control schemes are proposed in [20] to reduce the fluctuations in the estimated bandwidth of the basic AIMD (Additive Increase - Multiplicative Decrease) congestion control algorithm The following equations are used to make the distinction between AIMD and other binomial congestion control algorithms:

$$I: \quad cwnd_{t+RTT} \leftarrow cwnd_t + (\beta_1 / cwnd_t^k), \quad \beta_1 > 0,$$
$$D: \quad cwnd_{t+RTT} \leftarrow cwnd_t - (\beta_2 * cwnd_t^l), \quad 0 < \beta_2 < 1,$$

where I refers to an increase in window as a result of one window of acknowledgments in a round-trip time (RTT), D refers to the decrease in the congestion window upon detection of congestion by the sender and β_1 and β_2 are constants. A binomial algorithm is TCP-compatible if $k + l - 1$ and $l \leq 1$ for suitable β_1 and β_2 [20]. The choice of $\beta_1 = 1$ and $\beta_2 = 0.5$ is common practice. TCP-compatible algorithms are known to interact well with TCP and maintain the stability of the Internet. For $k = 0$ and $l = 1$, we obtain AIMD. The $k = 1$ and $l = 0$ case gives the IIAD (Inverse Increase-Additive Decrease) and the $k = 1/2$ and $l = 1/2$ case reduces to the SQRT algorithms which are shown to interact well with TCP AIMD across a wide range of network conditions over a RED bottleneck gateway (see [20] for a detailed discussion of TCP-compatible protocols). Figure 3 is devoted to the behavior comparison of AIMD and IIAD for the case where TCP/FACK with RPLD is used in a network with no ECN support coupled with the proposed selective discard algorithm (with $\alpha = 0.1$, $T_p = 5sec$) in order to transmit the stored video given in Section 5. Larger oscillations of *cwnd* (i.e., rate) in AIMD are reduced in IIAD at the expense of slow responsiveness to congestion. While AIMD is reducing its *cwnd* multiplicatively by $cwnd * \beta_2$ in case of a loss, IIAD decrements the rate by β_2, independent of the *cwnd* value. We leave a further comparison of binomial congestion control algorithms for future research and we'll use the basic AIMD parameters for the remainder of this paper.

5. EXPERIMENTAL RESULTS

In this section, we run two sets of simulations to show the performance of the proposed architecture in streaming video. The first set demonstrates the effect of TCP add-ons and selective frame discarding on the performance and the second set demonstrates the impact of T_p. For the rest of the section, the following notation will be used.

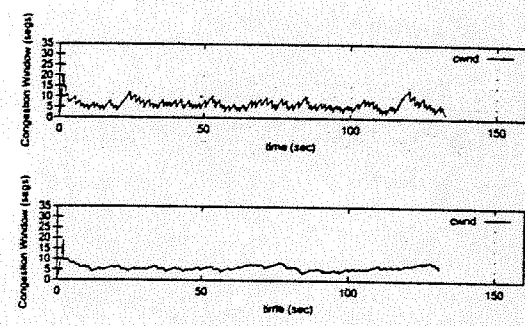
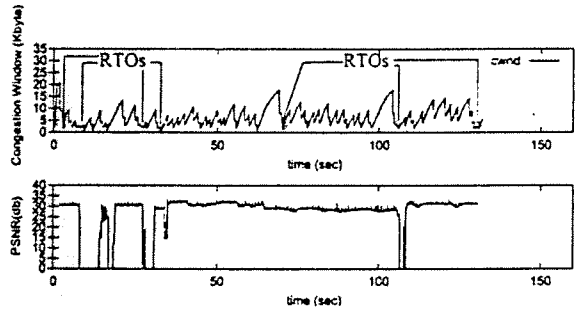


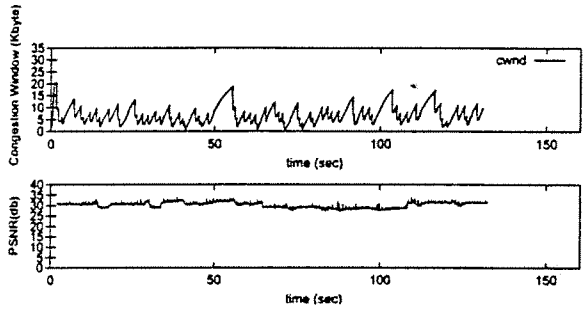
Fig. 3. For TCP/FACK source, AIMD ($k = 0.0, l = 1.0$) for top, IIAD ($k = 1.0, l = 0.0$) for bottom plot

N	: Number of video sources
T_p	: Initial receiver buffering in seconds.
C	: Bottleneck bandwidth
α	: Parameter that controls the selective discard algorithm
T_v	: Total video clip duration in seconds.
N_L, N_H	: Total number of low (L) and high priority (H) data in terms of Transport Layer(TL) packets.
$L_L^{(d)}, L_H^{(d)}$	: Low and high priority data discarded by the sender in terms of Transport Layer(TL) packets.
$L_L^{(p)}, L_H^{(p)}$	: Low and high priority data whether discarded by the sender or misses its playout time, in terms of Transport Layer(TL) packets. For finding packets that only miss the playout time, use $L_{L,H}^{(p)} - L_{L,H}^{(d)}$.
R_L, R_H	: Average bitrate of the total low and high priority frames generated by the encoder.
R'_L, R'_H	: Average bitrate of successfully played low priority and high priority frames

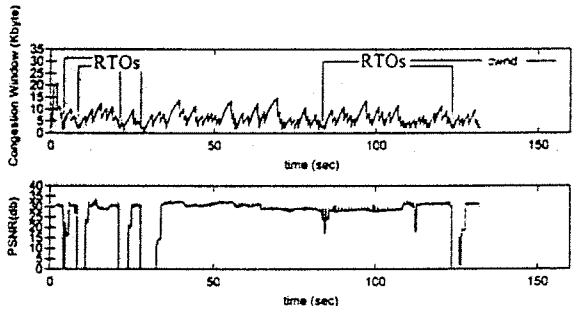
We used a modified version of ns-2 [22] as for the simulator to make the experiments. In all of the experiments, $N = 10$ video sources each of which is 130 secods long are active and packets from these sources arrive at a RED router sharing a single bottleneck link with capacity C. Each video source uses the TCP protocol with the same set of parameters and options while streaming the same video clip. There is one tagged source we monitor among the N sources particularly for PSNR (Peak Signal-Noise Ratio) plots. Each source starts streaming at random points in the video clip in order to prevent synchronization. Throughout the simulations, $R_L = 57917.78$ bps and $R_H = 51625.78$ bps corresponding to $N_L = 2286$ and $N_H = 1395$. RED parameters of the bottleneck router are given with the triple (minthresh, maxthresh, drop probability) = (20, 80, 0.25), where the router's physical queue length is set to 100 packets [21]. A comparative study is first carried out under the scenario of $C = 800$ kbps bottleneck link with selective frame discard with parameters $\alpha = 0.1$, and with $T_p = 2$



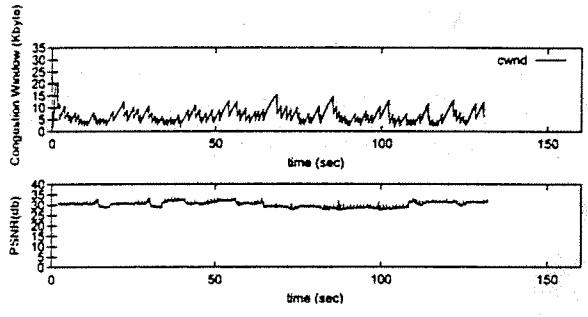
(a) TCP/Reno (no RPLD, no ECN)



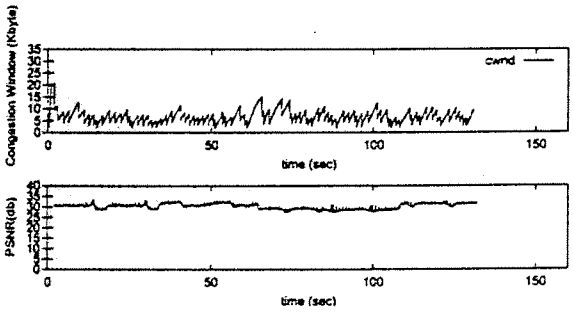
(b) TCP/Reno (no RPLD, with ECN)



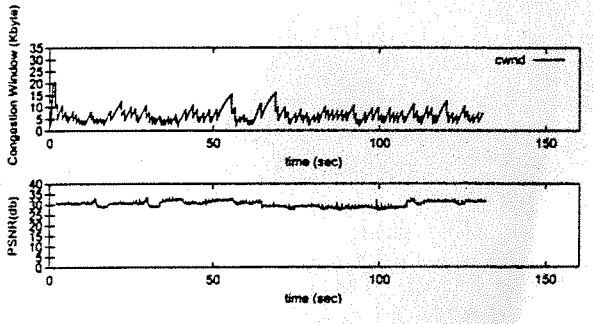
(c) TCP/FACK (no RPLD, no ECN)



(d) TCP/FACK (no RPLD, with ECN)



(e) TCP/FACK (with RPLD, no ECN)



(f) TCP/FACK (with RPLD, with ECN)

Fig. 4. Using 10 TCP sources with (k,l)=(0.0, 1.0) AIMD parameters, C=800 kbps, ($T_p = 2$ sec, $\alpha = 0.1$)

sec. initial buffering at the receiver. Figure 4 depicts the congestion window at the sender and the PSNR monitored at the receiver for six different transport protocol options. In (a) and (b), all sources use TCP/Reno but in (b) ECN is also enabled. We note that by using ECN, number of RTOs can substantially be eliminated in the standard TCP/Reno sender. This decrease in RTOs, results in an increase in the number of frames that reach their destination before their playout times, which then improves upon the observed PSNR. In Figure 4(c), TCP/FACK sender does not use any kind of data-driven loss recovery enhancement. Therefore, a number of RTOs occur which in turn cause many frames to miss their playout times, and consequently frequent drops in PSNR. However, when one or both of the loss recovery methods in Table 3 is applied, results in Figure 4(d,e,f) are obtained. No RTO is observed for these three cases at least in a 130 second video clip. In conclusion, improved PSNR results are obtained in Figure 4 by using the TCP add-ons provided in Table 3.

We present our quantitative results in Table 4, which are obtained by taking ensemble averages over ten simulations for a range of problem parameters (i.e., C and α are varied) and for a selected set of TCP variants, namely (a) TCP/Reno, (b) TCP/Reno with ECN, (c) TCP/FACK with RPLD and no-ECN. Table 4 is obtained for the case $T_p = 5$ sec., C is varied between 600 and 950 Kbps., and α ranges from 0.1 to 0.9. By observing the column for $L_H^{(p)}$, the number of high priority packets that miss their playout times can be obtained. Clearly, a high $L_H^{(p)}$ indicates an unacceptable video streaming performance. TCP/Reno with ECN and TCP/FACK with RPLD both improve upon the basic TCP/Reno but this improvement is more significant with the latter. As explained before, this improvement should be due to the reduction in the number of RTOs throughout the simulation.

Revisiting Table 4, we show that for $T_p = 5$ (i.e., initial buffering time is long enough), a conservative selective frame discard strategy (i.e., $\alpha = 0.1$) provides a robust video streaming performance over a wide range of scenarios. Aggressive frame discard strategies (i.e., $\alpha = 0.9$) tend to perform better in terms of the number of low priority frames injected into the network but at the expense of jeopardizing the successful playout of the high priority frames. This observation becomes critical especially in the case of low bottleneck bandwidth scenarios. We also observe that sources receive their fair share of the bandwidth (i.e., C/N) on the average. This can be checked by comparing $R'_L + R'_H$ and C/N in Table 4.

Next, we study the impact of T_p on the video streaming performance. We use TCP/FACK (with-RPLD) with the standard AIMD parameters ($k = 0.0, l = 1.0$) as the transport protocol, in a network with no ECN support. We use the relatively conservative frame discard strategy $\alpha = 0.1$.

Table 4. All results are for AIMD ($k = 0.0$, $l = 1.0$) parameters and $T_p = 5$ sec. $N_L = 2286$, $N_H = 1395$, and $R_L = 57917.78bps$, $R_H = 51625.78bps$

(C, α)	R'_L, R'_H (bps)		$L_L^{(d)}, L_H^{(d)}$ (pkts)		$L_L^{(p)}, L_H^{(p)}$ (pkts)	
(600, 0.1)	9245.44	44246.97	1893.90	0.00	1894.10	166.10
(600, 0.4)	9047.12	37568.19	1901.00	0.00	1901.20	335.10
(600, 0.7)	1044.01	26191.41	1821.20	0.00	1822.70	643.00
(600, 0.9)	8110.18	7888.18	1858.20	0.00	1934.70	1170.10
(800, 0.1)	26339.54	50812.07	1169.00	0.00	1169.50	6.80
(800, 0.4)	29897.03	49267.81	1023.20	0.00	1023.90	35.90
(800, 0.7)	26939.61	43319.26	1138.60	0.00	1141.30	183.70
(800, 0.9)	27813.86	28257.02	4050.70	0.00	1110.30	593.50
(950, 0.1)	41563.24	51308.95	598.00	0.00	590.80	0.20
(950, 0.4)	42361.99	50833.20	568.80	0.00	569.70	2.30
(950, 0.7)	45590.63	48615.21	410.40	0.00	411.90	48.60
(950, 0.9)	44872.22	43310.70	398.40	0.00	427.70	189.40

(a) TCP/Reno (no ECN) ($k = 0.0, l = 1.0$)

(C, α)	R'_L, R'_H (bps)		$L_L^{(d)}, L_H^{(d)}$ (pkts)		$L_L^{(p)}, L_H^{(p)}$ (pkts)	
(600, 0.1)	8954.03	48256.96	1888.60	0.00	1888.80	65.50
(600, 0.4)	7251.66	38068.99	1953.30	0.00	1953.40	326.00
(600, 0.7)	9176.75	30600.37	1880.90	0.00	1881.30	530.90
(600, 0.9)	5685.09	5770.31	1933.50	0.00	2022.10	1224.80
(800, 0.1)	30136.55	51264.78	1039.90	0.00	1040.80	0.10
(800, 0.4)	29467.36	50463.78	1038.70	0.00	1039.20	9.10
(800, 0.7)	27938.82	47312.08	1081.10	0.00	1081.90	74.80
(800, 0.9)	27024.47	30041.28	1034.40	0.00	1112.80	540.20
(950, 0.1)	41991.03	51334.89	571.80	0.00	572.80	0.00
(950, 0.4)	44474.26	50932.83	492.00	0.00	492.80	0.20
(950, 0.7)	45486.40	49768.01	418.70	0.00	419.70	13.90
(950, 0.9)	44501.37	44075.23	402.40	0.00	431.40	161.70

(b) TCP/Reno (with ECN) ($k = 0.0, l = 1.0$)

(C, α)	R'_L, R'_H (bps)		$L_L^{(d)}, L_H^{(d)}$ (pkts)		$L_L^{(p)}, L_H^{(p)}$ (pkts)	
(600, 0.1)	8008.73	51077.43	1922.60	0.00	1922.60	1.00
(600, 0.4)	8574.87	48116.98	1856.60	0.00	1887.70	62.80
(600, 0.7)	9367.22	38520.89	1849.50	0.00	1851.30	309.80
(600, 0.9)	5663.81	5324.80	1827.20	0.00	2021.50	1237.90
(800, 0.1)	30818.30	51208.27	997.80	0.00	998.40	0.40
(800, 0.4)	30760.05	50389.34	984.60	0.00	985.10	0.50
(800, 0.7)	29027.80	46900.64	1014.10	0.00	1015.20	6.90
(800, 0.9)	27167.86	29422.76	960.50	0.00	1124.40	561.30
(950, 0.1)	43427.35	51307.13	517.90	0.00	518.90	0.00
(950, 0.4)	43596.78	50968.60	477.00	0.00	477.90	0.10
(950, 0.7)	46603.95	50521.98	376.60	0.00	377.60	0.20
(950, 0.9)	41875.75	41824.23	447.10	0.00	523.80	214.80

(c) TCP/FACK (with RPLD, no-ECN) ($k = 0.0, l = 1.0$)

In this experiment, the video sequence at hand is streamed six times, in order to obtain a longer simulation. Since the same video is used six times, average high priority (H) and low priority (L) bit rates do not change and therefore $R_L = 57917.78$ bps and $R_H = 51625.78$ bps. The video clip length then becomes $T_v = 780 (= 6 * 130)$ sec., with $N_L = 6 * 2286$ and $N_H = 6 * 1395$ packets. In Figure 5, we provide the congestion window, frame rate, and the PSNR for three different values of $T_p \in \{1, 2, 5\}$. The PSNR plots demonstrate that with longer initial buffering times, one is able to absorb the fluctuations arising due to the AIMD behavior. Our results also clearly show that with short buffering, a good streaming performance cannot be obtained even with TCP/FACK with RPLD.

6. CONCLUSIONS

TCP (Transmission Control Protocol) with its well established AIMD-based congestion control algorithm is the prevailing transport layer protocol for non-real time data in the current Internet. The extensive operational experience behind TCP has led us to study the use of this congestion control scheme at the transport layer coupled with a selective frame discard strategy for video streaming applications. We propose to use TCP/FACK with RPLD so as to be able to reduce the number of RTOs that could otherwise have detrimental effects on the played video. We show that if the initial buffering time is kept long enough (e.g., $T_p = 5$), an efficient use of available bandwidth and acceptable video streaming performance are both attainable with a conservative frame discard strategy. Our results demonstrate that with short playout buffering times, none of the proposed schemes can compensate for the fluctuations in the estimated bandwidth although the use of ECN and/or FACK with RPLD helps relative to the pure TCP/Reno case.

7. REFERENCES

[1] S. Floyd, "Congestion Control principles", RFC 2914, September 2000.

[2] S. Floyd, M. Handley, and E. Kohler, "Problem statement for DCCP", Internet draft draft-ietf-dccp-problem-00.txt, October 2002.

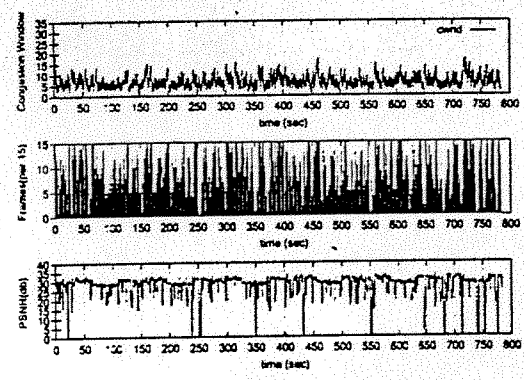
[3] S. Floyd, "TCP and Explicit Congestion Notification", ACM Computer Communication Review, V. 24 N. 5, pp. 10-23, October 1994.

[4] R. Rejaie, M. Handley, and D. Estrin, "Layered quality adaptation for Internet video streaming," IEEE Journal on Selected Areas in Communications, vol. 18, Dec. 2000.

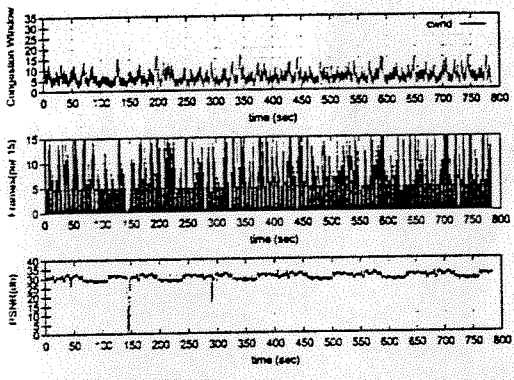
[5] P. H. Hsiao, H. T. Kung, K-S. Tan,"Video over TCP with receiver based delay control," Proc. 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video," Proc. NOSSDAV 2001, pp. 199-208.

[6] N. Feamster, D. Bansal, H. Balakrishnan, "On the Interactions Between Layered Quality Adaptation and Congestion Control for Streaming Video," Packet Video Worksop 2001.

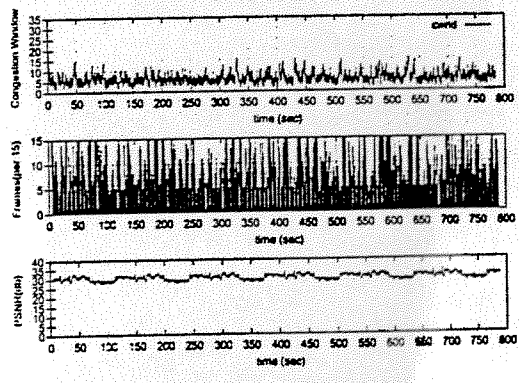
[7] A. Balan, O. Tickoo, I. Bajic, S. Kalyanaraman, J. Woods, "Integrated Buffer Management and Congestion Control for Video Streaming," GLOBECOM 2002.



(a) $T_p = 1$



(b) $T_p = 2$



(c) $T_p = 5$

Fig. 5. Using (k,l) = (0.0,1.0) parameters for 10 TCP/FACK(with-RPLD) sources, C=800 kbps, ($T_p = 1, 2, 5$ sec, $\alpha = 0.1$)

[8] H. Balakrishnan and S. Seshan, "The Congestion Manager," IETF, Nov.2000.

[9] D. Sisalem and H. Schulzrinne, "The Loss-Delay Adjustment Algorithm: A TCP-friendly Adaptation Scheme," Proc. NOSSDAV, July 1998.

[10] W. Feng, M.Liu, B. Krishnaswami, A. Prabhudev,"A Priority-Based Technique for the Delivery of Stored Video Across Best-Effort Networks," IS&T/SPIE Multimedia Computing and Networking 1999, San Jose, CA, Jan. 1999.

[11] D. Saparilla, K. Ross, "Optimal Streaming of Layered Video," INFOCOM 2000.

[12] G. Cote, B. Erol, M. Gallant, F. Kossentini, "H.263+: Video Coding at Low Bit Rates", IEEE Trans on Circuits and Systems for Video Tech, vol.8, no.7, November 1998.

[13] ITU-T Rec. H.263+, "Video Coding for Low Bit Rate Communication", 1998

[14] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[15] M.Allman, H.Balakrishnan, S.Floyd, "Enhancing TCP's Loss Recovery Using Limited Transmit", RFC3042, January 2001.

[16] M.Mathis, J.Mahdavi, S.Floyd, A.Romanov, "TCP Selective Acknowledgement Options", RFC2018, October 1996.

[17] M.Mathis, J.Mahdavi, "Forward Acknowledgement: Refining TCP Congestion Control", ACM SIGCOMM 1996.

[18] W. Stevens, "TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms", RFC 2001, January 1997.

[19] N.K.G. Samaraweera and G. Fairhurst, "Reinforcement of TCP Error Recovery for Wireless Communication", ACM SIGCOMM Computer Communication Review, Volume 28, Number 2 (April 1998).

[20] D.Bansal, H. Balakrishnan, "Binomial Congestion Control Algorithms", INFOCOM'2001.

[21] S. Floyd and V. Jacobson, "Random Early Detection gateways for Congestion Avoidance", IEEE/ACM Transactions on Networking, V.1 N.4, August 1993.

[22] "The Network Simulator - ns-2", http://www.isi.edu/nsnam/ns/

Face Verification Competition on the XM2VTS Database

Kieron Messer[1], Josef Kittler[1], Mohammad Sadeghi[1], Sebastien Marcel[2], Christine Marcel[2], Samy Bengio[2], F. Cardinaux[2], C. Sanderson[2], J. Czyz[3], L. Vandendorpe[3], Sanun Srisuk[4], Maria Petrou[1], Werasak Kurutach[4], Alexander Kadyrov[1], Roberto Paredes[5], B. Kepenekci[6], F. B. Tek[6], G. B. Akar[6], Farzin Deravi[7], and Nick Mavity[7]

[1] University of Surrey
Guildford, Surrey, GU2 7XH, UK
[2] Dalle Molle Institute for Perceptual Aritificial Intelligence
CP 592, rue du Simplon 4, 1920 Martigny, Switzerland
[3] Universite Catholique de Louvain, Batiment Stevin
Place du Levant 2, 1348 Louvain-la-Neuve, Belgium
[4] Mahanakorn University of Technology
51 Cheum-Sampan Rd, Nong Chok, Bangkok, 10530 Thailand
[5] DSIC, Universidad Politcnica de Valencia
Camino de Vera, s/n. 46022, Valencia, Spain
[6] Tübitak Bilten, ODTÜ Campus
0653, Ankara, Turkey
[7] Electronic Engineering Laboratory, University of Kent
Canterbury, CT2 7NT, UK

Abstract. In the year 2000 a competition was organised to collect face verification results on an identical, publicly available data set using a standard evaluation protocol. The database used was the Xm2vts database along with the Lausanne protocol [ ]. Four different institutions submitted results on the database which were subsequently published in [ ]. Three years later, a second contest using the same dataset and protocol was organised as part of AVBPA 2003. This time round seven seperate institutions submitted results to the competition. This paper presents the results of the competition and shows that verification results on this protocol have increased in performance by a factor of 3.

1 Introduction

In recent years the cost and size of biometric sensors and processing engines has fallen, a growing trend towards e-commerce, teleworking and e-banking has emerged and people's attitude to security since September 11th has shifted. For these reasons there has been a rapid increase in the use of biometric technology in a range of different applications. Many of these systems are based on the analysis of face images as they are non-intrusive and user-friendly. Moreover, personal identity can be ascertained without the client's assistance.

J. Kittler and M.S. Nixon (Eds.): AVBPA 2003, LNCS 2688, pp. 964–974, 2003.

However, face recognition technology is still developing and many papers on new face verification and recognition algorithms are being published almost daily. However, direct comparison of the reported methods can be difficult because tests are performed on different data with large variations in test and model database sizes, sensors, viewing conditions, illumination and background. Typically, it is unclear which methods are the best and for which scenarios they should be used. Evaluation protocols can help alleviate this problem.

Typically, an evaluation protocol defines a set of data, how it should be used by a system to perform a set of experiments and how the performance of the system should be quantified [16]. The protocol should be designed in such a manner that no bias in the performance is introduced. e.g. the training data is not used for testing. It should also represent a realistic operating scenario as different scenarios normally require different protocols, no single protocol will be able to cover all scenarios.

Over the past few years standard datasets for testing face authentication systems have become available. e.g. Yale [24], Harvard [21], Olivetti [23], M2VTS [22], ([1] gives a more comprehensive list). However, for many of them no associated protocol has been defined. Experiments carried out by different organisations on these datasets will divide the data into different test and training sets and consequentially they measure performance differently.

The FERET database has defined a protocol for face identification and face verification [18]. However, only a development set of images from the database are released to researchers. The remaining are sequestered by the organisers to allow independent testing of the algorithms. To date three evaluations have taken place, the last one in the year 2000 [17].

More recently, two Face Recognition Vendor Tests [2] have been carried out, the first in 2000 and the second in 2002. The tests are done under supervision and have time restrictions placed on how quickly the algorithms should compute the results. They are aimed more at indepently testing the performance of commercially available systems, however academic institutions are also able to take part. In the more recent test 10 commercial systems were evaluated.

In the year 2000 a competition on the Xm2vts database along with the Lausanne protocol [14] was carried out. Four different institutions submitted results on the database which were subsequently published in [15]. This paper presents the results of a second contest using the same dataset and protocol, that has been organised as part of AVBPA 2003. This time round seven seperate institutions submitted results to the competition.

The results published are based completely on self-assessment of the submitted methods by the participating research groups. All the data from the Xm2vts database to perform the tests is available from [9]. We believe that this open approach will increase, in the long term, the number of algorithms that will be tested on the XM2VTS database as each research institution is able to assess their algorithms performance at any time. To date over 100 institutions have obtained copies of the XM2VTS database.

The rest of this paper is organised as follows. In the next section the database and evaluation protocol are described. In section 3 an overview of each algorithm which entered the competition is given. In section 4 the results according to the protocol are presented along with a discussion. Finally, some conclusions are made.

2 The XM2VTS Database

The XM2VTS database [ ] is a multi-modal database consisting of face images, video sequences and speech recordings taken of 295 subjects at one month intervals. This database is available at the cost of distribution from the University of Surrey (see [ ] for details). The database is primarily intended for research and development of personal identity verification systems where it is reasonable to assume that the client will be cooperative. Since the data acquisition was distributed over a long period of time, significant variability of appearance of clients, e.g. changes of hair style, facial hair, shape and presence or absence of glasses, is present in the recordings - see figure 1.

The subjects were volunteers, mainly employees and PhD students at the University of Surrey of both sexes and many ethnical origins. The XM2VTS database contains 4 sessions. During each session two head rotation and "speaking" shots were taken. From the "speaking" shot, where subjects are looking just

Fig. 1. Sample images from XM2VTS database

below the camera while reading a phonetically balanced sentence, a single image with a closed mouth was chosen. Two shots at each session, with and without glasses, were acquired for people regularly wearing glasses.

For the task of personal verification, a standard protocol for performance assessment has been defined. The so called Lausanne protocol splits randomly all subjects into a client and impostor groups. The client group contains 200 subjects, the impostor group is divided into 25 evaluation impostors and 70 test impostors. Eight images from 4 sessions are used.

From these sets consisting of face images, training set, evaluation set and test set are built. There exist two configurations that differ by a selection of particular shots of people into the training, evaluation and test sets. The training set is used to construct client models. The evaluation set is selected to produce client and impostor access scores, which are used to find a threshold that determines if a person is accepted or not (it can be a client-specific threshold or global threshold). According to the Lausanne protocol the threshold is set to satisfy certain performance levels (error rates) on the evaluation set. Finally the test set is selected to simulate realistic authentication tests where impostor's identity is unknown to the system. The evaluation set is also used in fusion experiments (classifier combination) for training, but this is not relevant in the context of this paper.

The performance measures of a verification system are the False Acceptance rate (FA) and the False Rejection rate (FR). False acceptance is the case where an impostor, claiming the identity of a client, is accepted. False rejection is the case where a client, claiming his true identity, is rejected. FA and FR are given by:

$$FA = EI/I * 100\% \quad FR = EC/C * 100\% \tag{1}$$

where EC is the number of impostor acceptances, I is the number of impostor claims, EC the number of client rejections, and C the number of client claims. Both FA and an FR are influenced by an acceptance threshold. To simulate real application the threshold is set on the data from the evaluation set to obtain certain false acceptance (FAE) and false rejection error (FRE). The same threshold is afterwards applied to the test data and FA and FR on the test set are computed. Three thresholds are defined on the evaluation set:

$$\begin{aligned} T_{FAE=0} &= \arg\min_T(FRE|FAE = 0) \\ T_{FAE=FRE} &= (T|FAE = FRE) \\ T_{FRE=0} &= \arg\min_T(FAE|FRE = 0) \end{aligned} \tag{2}$$

Consequently, performance on the test set is characterised by six error rates:

$$\begin{matrix} FA_{FAE=0} & FR_{FAE=0} \\ FA_{FAE=FRE} & FR_{FAE=FRE} \\ FA_{FRE=0} & FR_{FRE=0} \end{matrix} \tag{3}$$

3 Overview of the Algorithms and the Scope of Their Evaluation

This section describes the face verification methods that participated in the contest. For this competition, it was decided just to report the results at the equal error rate, i.e. $FAE = FRE$.

Both configurations of the protocol are considered under two face image registration conditions: manual registration and fully automatic registration. Manual registration is self-explanatory. Fully automatic registration requires that the face has to be localised automatically for the test phase.

3.1 Best Results from ICPR2000 (Unis-ICPR2000)

In the ICPR 2000 competition, [13], the best verification results for both semi-automatic and fully automatic registration techniques were performed by a method developed at the University of Surrey. It was based on a technique reported in [16] which performs face verification based on linear discriminant analysis. A novel way of measuring the distance between probe image and the client template was used. We have included the results of this technique in this paper to give a baseline comparison and indicate how the algorithms have improved over the past three years.

3.2 Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)

IDIAP entered two seperate face verification algorithms into the competition. A brief description of each technique is given below.

IDIAP – Cardinaux The proposed face verification method is based on Gaussian Mixture Models (GMMs), [10] and [4]. The face images are analyzed on a block by block basis. Each block is decomposed in terms of an extension of the 2D Discrete Cosine Transform (DCT), namely DCT-mod2. The GMM approach uses a combination of Maximum Likelihood (ML) and Maximum a Posteriori (MAP) criteria.

IDIAP – Marcel We use skin color information in addition to the gray-level face image in order to train face verification systems using artificial neural networks, [12] and [11].

The representation used to code input images is based on gray-scale face image. The face bounding box is computed using manually located eyes coordinates. The face is cropped and the extracted sub-image is down-sized to a 30x40 image. After enhancement and smoothing, the face image becomes a feature vector of dimension 1200. The skin color feature is chosen to be simply the RGB color distribution of filtered skin pixels inside the face bounding box. For each

color channel, an histogram is built using 32 discrete bins. Hence, the feature vector produced by the concatenation of the 3 histograms (R, G and B) has 96 components.

Our face verification method is based on Multi-Layer Perceptrons (MLPs). For each client, an MLP is trained to classify an input to be either the given client or not. The input of the MLP is a feature vector corresponding to the face image with its skin color. The output of the MLP is either 1 (if the input corresponds to a client) or -1 (if the input corresponds to an impostor). The MLP is trained using both client images and impostor images, often taken to be the images corresponding to other available clients.

3.3 Universidad Politécnica de Valencia (UPV)

The local feature representation approach is used in this face verification contest, [15] [7]. Using this local feature representation scheme each image is represented by several smaller images. To classify each test image a nearest neighbor classifier is used by taking a suitable voting scheme. Given a test image, the k-nearest neighbors of its local feature images are found among the feature vectors computed for the training images. Each neighbor votes for its own class and a vector of votes (per class) is obtained by simply counting all votes. Following a direct voting scheme, the test image is classified into the most voted class. This sum rule of the votes of each local feature image is similar to the sum rule used in the Combining Classifiers theory.

3.4 Tübitak Bylten (TB)

The method uses a full Gabor wavelet transform for both finding feature points and extracting feature vectors [ ]. The feature extraction algorithm of the proposed method has two steps: (1) Feature point localization, (2). Feature vector generation. Feature vectors are extracted at points with high information content on the face image. The features are not limited to eyes, nose, etc., i.e. special facial features such as dimples are also extracted. The face image is then convolved with Gabor filters, and R_j is found to be the response of the face image to the jth Gabor filter. Feature localization is done by searching local maximums of R_j which are also having the value above the mean of all pixel values of R_j. Feature vectors are generated at the feature points as a composition of Gabor wavelet transform coefficients. To measure the similarity of two complex valued feature vectors, a normalized cross-correlation function is used which ignores the phase.

Face comparison is done in two steps. In the first step, the feature vectors of reference images those are not close enough to the feature vectors of the test image in means of both location and similarity, are eliminated. In the second step, the similarity of two faces is calculated as the mean of similarities of matched features.

3.5 Universite Catholique de Louvain (UCL)

This fuses results from three different face verification experts. It combines the 3 scores given by the algorithm using a weighted averaging. The first algorithm uses Gradient Direction Metric in the LDA subspace to compute the score (developed in UniS). The second algorithm uses the Probabilistic Matching to compute the score (developed in UCL). The third method computes the score by taking the L1 norm between the colour histogram of the face image (developed in UCL). The images are registered using manually located eye coordinates. More details can be found in [5].

3.6 Commercial System

The University of Kent used a well known commercial system to perform face verification using fully automatic registration according to the Lausanne protocol. The package was used with the default settings. In enrollment some images were rejected by the system. This meant that some client templates were built with only one or two examples. The package recommends a minimum of four suitable training images.

3.7 University of Surrey (UniS)

UniS entered three seperate face verification algorithms into the competition. A brief description of each is given below. The third method based on the the Shape Trace Transform was done in conjunction with a visiting researcher from the Mahanakorn University of Technology (MUT).

Normalised Correlation in LDA Space (UniS-NC) Linear Discriminant Analysis (LDA) projects the input image data into fisher faces which maximise the class separability. In [ ], it has been demonstrated that in the context of face verification, a matching score based on Normalised Correlation (NC) works effectively in the LDA space. Histogram equalisation was used to normalise the registered face photometrically. The thresholds in the decision making system have been determined using the Client-Specific Thresholding technique.

Error Correcting Codes (UniS-ECOC) In [5] a novel approach to face verification based on the Error Correcting Output Coding (ECOC) classifier was presented. In the training phase the client set is repeatedly divided into two ECOC specified subsets to train a set of binary classifiers. The output of the classifiers defines the ECOC feature space, in which it is easier to separate transformed patterns representing clients and impostors. The faces were first transformed in LDA space and the binary classifiers used to generate the binary codes were neural networks.

Shape Trace Transform (MUT-UniS-STT) A new face representation, the Shape Trace Transform (STT), for recognizing faces in an authentication system [?] has been developed. The STT offers an alternative representation for faces that has a very high discriminatory power. We estimate the dissimilarity between two shapes by a new measure we propose, the Hausdorff context. The reinforcement learning is used to search the optimal parameters of the algorithm, for which the within-class variance of the STT is minimized. This research demonstrates that the proposed method provides a new way for face representation. Our system is verified with experiments on the XM2VTS database.

4 Results and Discussion

Tables 1 and 2 shows the results using manual registration for both configurations I and II. The results on configuration I show that the best performing algorithm, the Shape Trace Transform, achieves an error rate of 1.47%. In fact three different methods have achieved a very similar low error rate, i.e. MUT-UniS-STT, UniS-NC and UniS-ECOC. This is an increase in performance by a factor of 3 over the best performing semi-automatic technique in the year 2000 competition where the best TER obtained was 4.8%.

Tables 3 and 4 shows the results using fully automatic registration for both configurations I and II. In the year 2000 competition the best performance for configuration I was 13.1%, in this competition it was 3.86%. An increase in performance of factor 3.5. Again, three different methods have achieved a similar level of performance, i.e. UPV, IDIAP-Cardinaux and UniS-NC.

Table 1. Error rates according to Lausanne protocol for configuration I with manual registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
UniS-ICPR2000	-	-	5.00	2.30	2.50	4. 80
IDIAP-Marcel	1.67	1.67	3.34	1.748	2.000	3.75
IDIAP-Cardinaux	0.75	2.00	2.75	1.84	1.50	3.34
MUT-UniS-STT	1.16	1.05	2.21	0.97	0.50	1.47
UCL	1.17	1.17	2.34	1.71	1.50	3.21
TB	2.34	1.00	3.34	5.61	5.75	11.36
UniS-ECOC	0.0	0.0	0.0	0.86	0.75	1.61
UniS-NC	0.33	1.33	1.36	0.48	1.00	1.48

Table 2. Error rates according to Lausanne protocol for configuration II with manual registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
IDIAP-Marcel	1.25	1.25	2.5	1.465	2.250	3.715
IDIAP-Cardinaux	0.75	0.75	1.50	1.04	0.25	1.29
TB	1.10	0.50	1.60	3.22	4.50	7.72
UniS-NC	0.33	0.75	1.08	0.25	0.50	0.75

Table 3. Error rates according to Lausanne protocol for configuration I using full automatic registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
UniS-ICPR2000	-	-	14.0	5.8	7.3	13.1
Commercial System	11.00	11.10	22.10	2.83	13.50	16.33
IDIAP-Cardinaux	1.21	2.00	3.21	1.95	2.75	4.70
UPV	1.33	1.33	2.66	1.23	2.75	3.98
UniS-NC	0.82	4.16	4.98	1.36	2.5	3.86

Table 4. Error rates according to Lausanne protocol for configuration II using full automatic registration

	Evaluation Set			Test Set		
Method	FA	FR	TER	FA	FR	TER
Commercial System	13.20	13.40	26.6	14.30	11.25	25.55
IDIAP-Cardinaux	1.25	1.20	2.45	1.35	0.75	2.10
UPV	1.75	1.75	3.50	1.55	0.75	2.30
UniS-NC	0.63	2.25	2.88	1.36	2.0	3.36

5 Conclusions

This paper presents a comparison of face verification algorithms that was organised in conjunction with the Audio Visual Biometric Person Authentication conference of 2003. Many different verification algorithms from 7 different institutions were tested using identical data from a large, publicly available multi-modal database, the XM2VTS. Training and evaluation was carried out according to an a priori known protocol. Results indicate that in the last three years the performance of the algorithms have increased by a factor of three.

References

[1] *The Face Recognition Homepage; http://www.cs.rug.nl/ peterkr/FACE/face.html.* 965

[2] *Face Recognition Vendor Tests: http://www.frvt.org.* 965

[3] *The XM2VTSDB: http://www.ee.surrey.ac.uk/ Research/VSSP/xm2vtsdb/.* 965, 966

[4] Fabien Cardinaux, Conrad Sanderson, and Sébastien Marcel. Comparison of mlp and gmm classifiers for face verification on xm2vts. In *To appear in the Proceedings of the Audio Visual Biometric Person Authentification*, Guildford, Surrey, June 2003. 968

[5] J. Czyz, J. Kittler, and L. Vandendorpe. Combining face verification algorithm. In R Kasturi, D Laurendeau, and C Suen, editors, *Proceedings 16th International Conference on Pattern Recognition III*, 2002. 970

[6] B. Kepenekci, F. B. Tek, and G. B. Akar. Occluded face recognition based on gabor wavelets. In *Proc International Conference on Image Processing*, September 2002. 969

[7] D. Keysers, R. Paredes, H. Ney, and E. Vidal. Combination of tangent vectors and local representations for handwritten digit recognition. In *International Workshop on Statistical Pattern Recognition*, 2002. 969

[8] J. Kittler, R. Gadheri, T. Windeatt, and J. Matas. Face verification via ecoc. In *Proceedings of British Machine Vision Conference 2001*, pages 593–602, 2001. 970

[9] J. Kittler, Y. P. Li, and J. Matas. On matching scores for lda-based face verification. In M Mirmehdi and B Thomas, editors, *Proceedings of British Machine Vision Conference 2000*, pages 42–51, 2000. 970

[10] Y. P. Li, J. Kittler, and J. Matas. On Matching Scores of LDA-based Face Verification. In Tony Pridmore and Dave Elliman, editors, *Proc British Machine Vision Conference BMVC2000*, page submitted, London, UK, September 2000. University of Bristol. British Machine Vision Association. 968

[11] Sébastien Marcel and Samy Bengio. Improving face verification using skin color information. In *Proceedings of the 16th International Conference on Pattern Recognition*. IEEE Computer Society Press, 2002. 968

[12] Sébastien Marcel, Christine Marcel, and Samy Bengio. A state-of-the-art Neural Network for robust face verification. In *Proceedings of the COST275 Workshop on The Advent of Biometrics on the Internet*. Rome, Italy, 2002. 968

[13] J. Matas, M. Hamouz, K. Jonsson, J Kittler, Y. P Li, C Kotropoulos, A Tefas, I Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Comparison and face verification results on the xm2vts database. In A Sanfeliu, J J Villanueva, M Vanrell, R Alquezar, J Crowley, and Y Shirai, editors, *Proceedings of International Conference on Pattern Recognition, Volume 4*, pages 858–863, 2000. 964, 965, 968

[14] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999. 964, 965, 966

[15] R. Paredes, J. C. Pérez, A. Juan, and E. Vidal. Local representations and a direct voting scheme for face recognition. In *In Proc. of the Workshop on Pattern Recognition in Information Systems*, July 2001. 969

[16] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki. An introduction to evaluating biometric systems. *IEEE Computer*, pages 56–63, February 2000. 965

[17] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. volume 22, pages 1090–1104, October 2000. 965

[18] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The FERET database and evaluation procedure for face-recognition algorithm. *Image and Vision Computing*, 16:295–306, 1998. 965

[19] C. Sanderson. *Automatic Person Verification Using Speech and Face Information.* PhD thesis, Griffith University, Brisbane, Australia., 2002. 968

[20] Sanun Srisuk, Maria Petrou, Werasak Kurutach, and Alexander Kadyrov. Face authentication using the trace transform. In *To appear in CVPR2003*. IEEE Computer Society Press, 2003. 971

[21] `ftp://hrl.harvard.edu/pub/faces`. 965

[22] `http://nsl.tele.ucl.ac.be/M2VTS/`. 965

[23] `http://www.cam-orl.co.uk/facedatabase.html`. 965

[24] `http://cvc.yale.edu/projects/yalefaces/yalefaces.html`. 965

BİBLİYOGRAFİK BİLGİ FORMU	
1- Proje No: **101E036**	2- Rapor Tarihi: **15. Mart. 2006**
3- Projenin Başlangıç ve Bitiş Tarihleri: **1. Ağustos. 2002 - 1. Ağustos. 2005**	
4- Projenin Adı: **COST 276 "Tümleşik çoklu ortam iletişimi için bilgi idaresi" (Information and Knowledge Management for Integrated Media Communication Systems)**	
5- Proje Yürütücüsü ve Yardımcı Araştırıcılar: Doç. Dr. Gözde Bozdağı Akar, Doç. Dr. Aydın Alatan	
6- Projenin Yürütüldüğü Kuruluş ve Adresi: ODTÜ, Elektrik Elektronik Müh. Blm., Ankara	
7- Destekleyen Kuruluş(ların) Adı ve Adresi: TÜBİTAK	
8- Öz (Abstract): Günümüzde çoğulortamlı teknolojilerin, iletişim ve yayıncılık sektörünün hızlı bir şekilde yakınsamasına şahit olmaktayız. Etkileşimli televizyon, ısmarlama video (video-on-demand), sayısal kayıt cihazları bu yakınsamanın ortaya çıkardığı önemli ürünler olarak bilinmektedir. Bu ürünlerin geliştirilmesi ve amaçlarına uygun hizmet edebilmesi için önemli olan unsurlar kullanıcı servislerinin verimli teslimi, otomatik içerik işleme, kişiselleştirme olarak sayılabilir. Bunların gerçekleştirilebilmesi için bu proje kapsamında 4 ana madde üzerinde yoğunlaşılmıştır : İçerik ve bilgi içeren öteveri (metadata) tanımları, Çoğulortamlı sistem teknolojileri , Damgalama, Çoğulortamlı veri yönetimi.	

Anahtar Kelimeler: Çoğulortam, veritabanı, gezgin iletişim, damgalama, veri yönetimi
9- Proje ile ilgili Yayın/Tebliğlerle ilgili Bilgiler Ekde verilmiştir.
10- Bilim Dalı Doçentlik B. Dalı Kodu: Telekomünikasyon ISIC Kodu: Uzmanlık Alanı Kodu:
11- Dağıtım(*): ☐ Sınırlı ☐ Sınırsız x
12- Raporun Gizlilik Durumu: ☐ Gizli ☐ Gizli Değil x